COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

PREDICTIVE MODELS ON DISASTER DECLARATIONS IN THE US

Autor: María Araujo Pérez
Director: Allison Coffey Reilly

Madrid
Julio de 2019

AUTHORIZATION FOR DIGITALIZATION, STORAGE AND DISSEMINATION IN THE NETWORK OF END-OF-DEGREE PROJECTS, MASTER PROJECTS, DISSERTATIONS OR BACHILLERATO REPORTS

1. Declaration of authorship and accreditation thereof.

The author Mr. /Ms. María Araujo Pé

HEREBY DECLARES that he/she owns the intellectual property rights regarding the piece of work: Predictive Models for Disaster Declarations in the US

that this is an original piece of work, and that he/she holds the status of author, in the sense granted by the Intellectual Property Law.

2. Subject matter and purpose of this assignment.

With the aim of disseminating the aforementioned piece of work as widely as possible using the University's Institutional Repository the author hereby **GRANTS** Comillas Pontifical University, on a royalty-free and non-exclusive basis, for the maximum legal term and with universal scope, the digitization, archiving, reproduction, distribution and public communication rights, including the right to make it electronically available, as described in the Intellectual Property Law. Transformation rights are assigned solely for the purposes described in a) of the following section.

3. Transfer and access terms

Without prejudice to the ownership of the work, which remains with its author, the transfer of rights covered by this license enables:

a) Transform it in order to adapt it to any technology suitable for sharing it online, as well as including metadata to register the piece of work and include "watermarks" or any other security or protection system.

b) Reproduce it in any digital medium in order to be included on an electronic database, including the right to reproduce and store the work on servers for the purposes of guaranteeing its security, maintaining it and preserving its format.

c) Communicate it, by default, by means of an institutional open archive, which has open and cost-free online access.

d) Any other way of access (restricted, embargoed, closed) shall be explicitly requested and requires that good cause be demonstrated.

e) Assign these pieces of work a Creative Commons license by default.

f) Assign these pieces of work a HANDLE (*persistent* URL). by default.

4. Copyright.

The author, as the owner of a piece of work, has the right to:

a) Have his/her name clearly identified by the University as the author

b) Communicate and publish the work in the version assigned and in other subsequent versions using any medium.

c) Request that the work be withdrawn from the repository for just cause.

d) Receive reliable communication of any claims third parties may make in relation to the work and, in particular, any claims relating to its intellectual property rights.

5. Duties of the author.

The author agrees to:

a) Guarantee that the commitment undertaken by means of this official document does not infringe any third party rights, regardless of whether they relate to industrial or intellectual property or any other type.

b) Guarantee that the content of the work does not infringe any third party honor, privacy or image rights.
c) Take responsibility for all claims and liability, including compensation for any damages, which may be brought against the University by third parties who believe that their rights and interests have been infringed by the assignment.
d) Take responsibility in the event that the institutions are found guilty of a rights infringement regarding the work subject to assignment.

6. Institutional Repository purposes and functioning.

The work shall be made available to the users so that they may use it in a fair and respectful way with regards to the copyright, according to the allowances given in the relevant legislation, and for study or research purposes, or any other legal use. With this aim in mind, the University undertakes the following duties and reserves the following powers:

a) The University shall inform the archive users of the permitted uses; however, it shall not guarantee or take any responsibility for any other subsequent ways the work may be used by users, which are non-compliant with the legislation in force. Any subsequent use, beyond private copying, shall require the source to be cited and authorship to be recognized, as well as the guarantee not to use it to gain commercial profit or carry out any derivative works.
b) The University shall not review the content of the works, which shall at all times fall under the exclusive responsibility of the author and it shall not be obligated to take part in lawsuits on behalf of the author in the event of any infringement of intellectual property rights deriving from storing and archiving the works. The author hereby waives any claim against the University due to any way the users may use the works that is not in keeping with the legislation in force.
c) The University shall adopt the necessary measures to safeguard the work in the future.
d) The University reserves the right to withdraw the work, after notifying the author, in sufficiently justified cases, or in the event of third party claims.

Madrid, on 11 of July, 2019

HEREBY ACCEPTS

Signed:

Reasons for requesting the restricted, closed or embargoed access to the work in the Institution's Repository

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Predictive Models on disaster declarations in the US

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso

académico 2018/2019 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada

de otros documentos está debidamente referenciada.

Fdo.: Maria Araujo Perez Fecha: 08/ 07/ 2019

DocuSigned by:
Maria Araujo
FA5ED35D9DDF47E...

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Allison Coffey Reilly Fecha: 7/8/2019

DocuSigned by:
Allison Coffey Reilly
01A6D14B1704445...

COMILLAS
UNIVERSIDAD PONTIFICIA

ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

PREDICTIVE MODELS ON DISASTER DECLARATIONS IN THE US

Autor: María Araujo Pérez
Director: Allison Coffey Reilly

Madrid
Julio de 2019

MODELOS DE PREDICCION DE CATASTROFES NATURALES EN LOS ESTADOS UNIDOS

Autor: Araujo Pérez, María

Director: Reilly, Allison C.

Entidad Colaboradora: ICAI-Universidad Pontifica Comillas

RESUMEN DEL PROYECTO

Introducción

Todos los años, frente a las catástrofes naturales que tienen lugar en los Estados Unidos, los gobiernos estatales y federales son los responsables de proporcionar asistencia a los ciudadanos, así como de llevar a cabo las obras de reconstrucción.

La Ley Stafford para la reforma de recuperación de desastres naturales y sus posteriores reformas fue diseñada para establecer un plan de acción para situaciones de emergencia, así como un un plan para coordinar y establecer quién debe ser el responsable de facilitar asistencia, el gobierno del Estado o el gobierno federal.

Cuando tiene lugar un desastre natural, el gobierno del condado y estado son los responsables de ejecutar el plan de emergencia. En ciertos casos, recogidos en la Ley Stafford, los gobiernos locales podrán solicitar al presidente del país que haga lo que se denomina como "Declaración de Catástrofe Natural", la cual va acompañada de ayudas financieras y materiales proporcionadas por la Agencia Federal para la Gestión de Emergencias (FEMA). Esta declaración de desastre natural se realiza cuando se cumplen ciertas circunstancias, como por ejemplo la magnitud de la catástrofe, los daños provocados, el número de personas afectadas.

El hecho de que un evento sea declarado como "Catástrofe Natural" tiene un gran impacto económico tanto en el gobierno del estado como en sus ciudadanos, de ahí la importancia de entender bien cuáles son las condiciones o características que determinan la obtención de dicha declaración.

La persona responsable en última instancia de decidir si un desastre alcanza la categoría de Catástrofe Natural es el presidente de los Estados Unidos. Estas decisiones pueden estar influenciadas por motivos políticos o de otra índole (por ejemplo, presión mediática). No existe un modelo objetivo y transparente para determinar si un desastre natural alcanza la categoría de Catástrofe natural.

El objetivo de este estudio es el de construir un modelo científico para predecir si un desastre natural va a conllevar una Declaración Presidencial de Catástrofe Natural en el estado de Maryland. El conjunto de datos utilizados para este proyecto incluye las Declaraciones de Catástrofes de los 15 últimos años (2003-2018) en el estado de Maryland por condados, así como otras variables como por ejemplo ingresos, población,

daños a los cultivos y a las propiedades, número de heridos y fallecimientos, y magnitud del desastre entre otras.

Este modelo permitiría mejorar el enfoque actual de cómo determinar si se debe hacer o no una Declaración Presidencial de Catástrofe Natural, una cuestión relevante e importante dada la alta incidencia de catástrofes naturales en EE. UU. y la gran cantidad de dinero que está en juego. Para poner en contexto, solo en 2017 el Gobierno de Estados Unidos se gastó $307 billones en asistencia en catástrofes.

Metodología

Para la elaboración de este proyecto se ha utilizado un modelo de regresión logístico binario, dado que la variable dependiente es una variable dicotómica (si la catástrofe es declarada o no). El modelo, junto a algunas variables independientes que serán explicadas en el estudio (tanto categóricas como continuas), cumplen con los dos primeros principios básicos de un modelo de regresión logístico binario.

Para la construcción de este modelo se han utilizado tres fuentes de datos: Storm Events Database de la Administración Nacional Oceánica y Atmosférica (NOAA), la Oficina del Censo de EEUU, y la universidad de Maryland. Tras la limpieza y consolidación de los datos, el resultado ha sido una gran base de datos que contiene toda la información sobre el tipo de eventos que sucedieron en el estado de Maryland del año 2003 al 2018, a nivel de condado, incluyendo si han sido declarados o no como Catástrofes por el Gobierno Federal, y otras variables económicas y no económicas (tamaño de la población, ingresos en el hogar, número de heridos y muertos, daños económicos en las cosechas y propiedades, entre otros).

Las variables independientes incluidas en el análisis son las siguiente: Ingresos, Población, Lesiones Directas, Lesiones Indirectas, Fallecimientos Directos, Fallecimientos Indirectos, Tipo de Evento, Daño a los Cultivos y Daño a la Propiedad como variables independientes; y la declaración de catástrofe natural como variable dependiente binaria.

El primer paso para construir el modelo es entender la relación de cada variable independiente con la dependiente. Con esta finalidad, los siguientes análisis fueron llevados a cabo: tablas de contingencia, test de chi-cuadrado, t-student y regresión logística univariable.

En los modelos de regresión logística, puesto que el modelo incluye logaritmos, el estudio del término beta no es constante, al contrario que ocurre en una regresión lineal. Por consiguiente, para estudiar el efecto constante de la variable independiente se utiliza el término de OR=(exp(betha)). Este último representa la posibilidad de que un determinado desastre natural sea declarado como catástrofe si se cumple una determinada condición con respecto a ese mismo desastre si no se cumple esa misma condición.

Para construir el modelo se han utilizado el método de *backward elimination.* Este método consiste en construir un primer modelo con todas las variables, para posteriormente ir eliminado variables una a una si no tienen importancia estadística. Los criterios utilizados para saber si una variable es significativa estadísticamente son el p-valor y el *test the Wald*.

Finalmente, el modelo se ha evaluado mediante la realización de los tests de bondad del ajuste: logaritmo de verosimilitud -2, Pseudo-R^2 de Cox y Snell y de Nagelkerke, especificidad y sensibilidad, y el test de Hosmer-Lemeshow.

Resultados

El efecto de la variable categórica "tipo de evento" es estudiada mediante un test de chi-cuadrado para evaluar su relación con la variable dependiente. El resultado del test es un valor X^2=123.375>11.07 con un p-valor de 0.05 que permite rechazar la hipótesis nula, indicando por lo tanto la dependencia de ambas variables.

Las variables continuas se estudian mediante un test de independencia *t-student* de comparación de medias, seguido de un estudio de regresión logística univariable. Los resultados muestran que las siguientes variables son respaldadas por los test, capaces de rechazar la hipótesis nula, y por consiguiente son aquellas capaces de demostrar la desigualdad de medias y la relación significativa entre variables: heridos directos, heridos indirectos, Fallecimientos Directos, Daños a los cultivos y Daño a la Propiedad.

El resto de variables (Ingresos, Población, Fallecimientos Indirectos) no muestran indicios de ser significantes estadísticamente para el futuro modelo.

Variable	**P-value**	**Odds-Ratio**	**Lower $CI_{95\%}$**	**Upper $CI_{95\%}$**
Ingresos	0.5322	1	1	1
Población	0.17	1	1	1
Tipo de Evento(1)-Granizo	0.071	1.962	0.943	4.079
Tipo de Evento (2)-Otros	0	4.710	2.664	8.328
Tipo de Evento (3)-Inundación repentina	0	9.480	5.193	17.306
Tipo de Evento (4) -Tormenta eléctrica	0	3.255	1.800	5.884
Tipo de Evento (5)-Tormenta de invierno	0	9.978	5.313	18.741
Heridos Directas	0.062	1.044	0.998	1.093
Heridos Indirectas	0.002	1.744	1.224	2.487
Fallecimientos Directos	0.0061	1.617	1.147	2.281
Fallecimientos Indirectos	0.0979	2.426	0.849	6.929
Daños a los cultivos	0.004	1	1	1
Daño a la probabilidad	0	1	1	1

Tabla 1: resumen del análisis bivariable para cada variable independiente

En la tabla de arriba se puedes observar los Odd-ratios desajustados. Es importante destacar que los valores para las variables continuas podrían estar sesgadas, puesto que el OR representa una diferencia de posibilidades para cada observación con respecto a la anterior. El valor es un simple número de referencia. Por esta razón en ocasiones es útil estudiar el OR categorizando las variables continuas para ver el efecto de cada grupo, a pesar de que el método de clasificación en estadística es muy complicado y se puede perder información por el camino.

Cuando se está construyendo el modelo logístico, uno podría utilizar el método de *forward selection*, que consiste en añadir variables consideradas significantes en base al análisis bivariable en cada paso. O podría utilizar el método de eliminación hacia atrás, que es el método elegido en este caso.

Hay dos formas en las que se ha aplicado el método de *backward elimination*: utilizando el p-valor o el test de Wald. Ambos coinciden en eliminar las mismas variables en el mismo orden, pero terminando en distintos puntos. El resultado de los parámetros más importantes se muestra en la tabla inferior.

Las variables eliminadas por orden de eliminación son las siguientes: Fallecimientos Indirectos, Heridos Directos, Fallecimientos Directos, Ingresos, Población y Daños a los Cultivos.

Model	Log-likelihood	Cox and Snells R^2	Nagelkerke R^2	Specificity	Sensitivity	Overall Percentage
1	2661.769	0.047	0.111	99.9	6.7	92.6
2	2661.830	0.047	0.111	99.9	6.7	92.6
3	2663.653	0.046	0.110	99.9	6.7	92.7
4	2665.257	0.046	0.110	99.9	6.5	92.6
5	2667.5	0.046	0.109	99.9	6.5	92.6
6	2670.22	0.045	0.107	99.9	6.5	92.6
7	2679.659	0.044	0.103	99.9	5.8	92.6

Tabla 2: resumen de los parámetros importantes de la bondad de ajuste para los siete modelos propuestos

Model	Chi-cuadrado	gl	Sig.
1	13.669	8	0.091
2	13.835	8	0.086
3	11.894	8	0.156
4	14.261	8	0.075
5	6.511	9	0.59
6	20.314	7	0.005
7	23.758	7	0.001

Tabla 3: resumen del test Hosmer-Lemeshow para los siete modelos propuestos

Si el p-valor es elegido como el criterio de decisión para eliminar variables, el modelo seleccionado es el número 7, en el cual las variables eliminadas por orden de eliminación son: Fallecimientos Indirectos, Heridos Directas, Fallecimientos Directos, Ingresos, Población y Daños a los Cultivos.

Por otro lado, empleando el método de *backward elimination* con el test de Wald automáticamente con el programa SPSS, el método se para en el modelo número 6, dónde las variables que acabo de mencionar son todas eliminadas menos Daños a los Cultivos.

Mirando a todos los parámetros en global, el modelo final elegido para el estudio es el modelo número 5. Este modelos es en el que el test de Hosmer_Lemeshow no es capaz de rechazar la hipótesis nula que dice que el modelo se ajusta a los datos; y a la vez, muestra los parámetros más aceptables para: el logaritmo de verosimilitud (cuanto más alto mejor), el rango descrito por las Pseudo-R^2 más amplio (el 4.6-10.9% de la variable dependiente es explicada por las variables independientes) y unos valores más altos de especificidad (99.9%) y sensibilidad (6.7%).

A pesar de ser el mejor modelo obtenido, no es un buen modelo. El rango mostrado por las Pseudo-R^2 es demasiado bajo (implica la posible ausencia de alguna variable independiente), la especificad es demasiado alta y la sensibilidad demasiado baja. Eso sin mencionar que el modelo incluye dos variables no estadísticamente significantes (Población con un p-valor de 0.119 y Daño a los cultivos con un p-valor de 0.099).

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Population	.000	.000	2.445	1	.118	1.000	1.000	1.000
	Injuries Indirect	.441	.199	4.938	1	.026	1.555	1.053	2.295
	Event Type			97.040	5	.000			
	Event Type(1)	.671	.374	3.222	1	.073	1.956	.940	4.068
	Event Type(2)	1.386	.292	22.476	1	.000	4.000	2.255	7.095
	Event Type(3)	2.181	.309	49.850	1	.000	8.857	4.834	16.228
	Event Type(4)	1.166	.303	14.821	1	.000	3.210	1.773	5.812
	Event Type(5)	2.208	.324	46.524	1	.000	9.100	4.824	17.163
	Dge_crop	.000	.000	2.714	1	.099	1.000	1.000	1.000
	Dge_prop	.000	.000	35.210	1	.000	1.000	1.000	1.000
	Constante	-3.864	.282	187.423	1	.000	.021		

a. Variables especificadas en el paso 1: Population, Injuries Indirect , Event Type, Dge_crop, Dge_prop.

Tabla 4: Modelo final elegido (nº5) de regresión logística binaria

En la tabla superior, se puede observar cómo entre en análisis bivariable y el modelo de regresión logística final, los valores de las razones de oportunidades se mantienen prácticamente constantes. Estas similitudes muestran relaciones fuertes entre las variables independientes del modelo y la dependiente.

Volviendo a los resultados desfavorables de la bondad de ajuste, la alta especificidad debe ser destacada por su alto valor. Esto significa que los casos negativos (las no declaraciones) se predijeron en un porcentaje muy elevado, especialmente si lo comparamos al número de casos positivos (las declaraciones) que se pueden predecir bien (baja sensibilidad).

En estadística resulta menos perjudicial cometer error de tipo II que de tipo I, parece una mejor opción predecir que un evento va a ser declarado catástrofe natural cuando no lo va a ser, a decir lo contrario. En vez de intentar reducir el umbral de probabilidad donde se considera catástrofe natural o no (por defecto es 0,5), se observan las curvas de Característica Operativa del Receptor (COR).

La curva de COR analizada la relación entre especificidad y sensibilidad. Todas las variables de las áreas bajo la curva oscilan alrededor del valor de 0,5, lo que significa que el modelo no es capaz de separar los grupos positivos de los negativos.

Este análisis confirma que nuestro modelo predice que muchos eventos no conseguirán la Declaración de Catástrofe Natural cuándo en realidad sí deberían. Este desequilibrio hace que el modelo no sea capaz de distinguir correctamente entre los diferentes casos, dando lugar a una predicción más débil, al contrario de lo que parecía al principio.

Conclusión

Los modelos logísticos son una herramienta muy útil cuando se trata de predecir una variable dicotómica. Debido al uso de logaritmos en el modelo, la interpretación se hace mediante el uso de las razones de oportunidad.

Mediante el método de *backward elimination* y el p-valor, se realizaron varias iteraciones del modelo con diferentes variables. El modelo con el poder predictivo más fuerte es el modelo que incluye las siguientes variables: tamaño de la población, heridas indirectas, tipo de evento, daños a los cultivos y daño a la propiedad. Este modelo incluye dos variables que no parecen ser estadísticamente significativas de acuerdo al p-valor. Estas variables son Población (con un p-valor=0.118) y Daño a los Cultivos (p-valor=0.099). A veces la mejor opción puede ser dejar alguna variable que no parece ser estadísticamente significativa en el modelo que simplemente añada información antes de perder fuerza predictiva en los otros parámetros analizados (logaritmo de verosimilitud, Pseudo-R^2, test de Hosmer-Lemeshow, especificidad y sensibilidad). En este caso, la diferencia es tan pequeña que dejar las dos variables es un riesgo asumible.

A pesar de que los parámetros elegidos son los más aceptables en este modelo, el parámetro de especificidad del 99,9% muestra las limitaciones del modelo en cuanto a mostrar un número de predicciones para aquellos casos positivos, siendo estos aquellos en los que el evento es declarado como catástrofe natural. La sensibilidad del 6,5% junto al estudio del área bajo la curva COR arrojó luz sobre el hecho de que la muestra no contenía suficientes casos positivos de Catástrofes naturales declaradas en comparación con todos los desastres naturales ocurridos (415 contra 4925, ~8%). Este es el origen del problema: no hay suficientes casos de Catástrofes naturales declarados en la base de datos utilizada.

Una posible solución al problema podría ser utilizar uno de los métodos proporcionados por *Machine Learning*, como árboles de decisión, sobremuestreo o submuestreo.

PREDICTIVE MODELS FOR DISASTER DECLARATIONS IN THE US

Introduction

In face of a natural disaster in the United States, both county, state and federal governments are responsible for aiding citizens financially and physically to recover from the damages. The Stafford Disaster Relief and Emergency Assistance Act and its later amendments was designed to bring a systematic approach to coordinate who and when needs to provide natural disaster assistance. When a natural disaster occurs, the government of the county and state in which the natural disaster takes place must execute the state's emergency plan. On top of county and state assistance, the state government can request the US government to make a "Presidential disaster declaration", which then triggers federal financial and physical assistance through the FEMA (Federal Emergency Management Agency). This declaration is made only when certain circumstances are met (impact of disaster, scale of disaster, to name a few).

Whether a natural disaster triggers a Presidential Disaster Declaration has a big economic and financial impact on both the State government and its citizens, and therefore it is very important to understand what are the conditions or characteristics that will lead to such declaration.

Presidential disaster declarations in the USA are very dependent on the judgement of the President who is the person ultimately in charge of making the decision on whether to take action and help a county or a state when a hazard occurs. Furthermore, there are other factors that influence, such as how and when the Governor requests and completes the form and the reasons he/she states to request such help.

The objective of this study is to build a scientific model to predict the likelihood that a given natural disaster would lead to a Presidential Disaster Declaration in the state of Maryland. The data set used in this project includes the Disaster Declarations of the past 15 years (2003-2018) in the state of Maryland by county, as well as several variables such as income, population, damage crops and properties, number of injuries and deaths, and size of the disaster among others.

This model would improve the existing approach of when to make a Presidential Disaster Declaration, a very relevant and important topic given the high occurrence of natural disasters in the US and the large amounts of money at stake.

Methodology

The predictive model will be built using a logistic regression, as the dependent variable that has to be explained is a binary decision (whether or not the disaster is declared). The model along with some independent variables that will be explained (either categorical or continuous) meet the first two basic assumptions of a logistic model.

To build this model, three different sources of data were used: the Storm Events database from the National Oceanic and Atmospheric Administration (NOAA), the US Census Bureau, and the university of Maryland. Cleaning and consolidating the data from the different sources. The result of cleaning and consolidating all data sources was a large database with all event types that occurred in the state of Maryland from 2003 to 2018, on a county level, including whether they had been declared a Disaster by the Federal government, and other economic and non-economic data by county (population size, household income and so on).

The independent variables included in the analysis were the following: Income, Population, Direct Injuries, Indirect Injuries, Indirect Deaths, Event Type, Damage Crops and Damage Property as independent variables and Disaster as the dependent variable.

The first step to build the model is to understand the relationship of each independent variable with the dependent one. For this purpose, the following tests were conducted: contingency tables, chi-tests, t-student and univariate logistic regression.

The next step is to look at the odds-ratio. In logistic regression models, the parameter of study betha is not constant, as it is the case in a linear regression model. Therefore, to study the constant effect of a variable, the odds-ratio must be studied OR=(exp(betha)). The Odds-ratio represents the odds of an event being declared a disaster in each case. Important to mention is that it measures the odds, not probability.

The model is then built by entering all the variables at first and eliminating one at a time based on the p-value criteria (measure of the statistical significative measure) or the Wald test (backward elimination).

Finally, we need to evaluate how good the model fits the data. For this purpose, we will conduct tests such as Log-likelihood, Pseudo-R^2, Specificity and Sensibility and Hosmer-Lemeshow.

Results

The categorical variable Event type effect is studied by the chi-squared test to check for its relationship to the independent variable Declaration. Getting a result of X^2=123.375>11.07 with a p-value of 0.05 allows the null hypotheses to be rejected thus stating the dependency of both variables.

The continuous variables are also studied but with an independent variable t-student test of means comparison, followed by a univariate logistic regression. The results showed that the tests able to reject the null hypotheses that stated equal means thus concluding that have a significant difference in values for each class type (declared/not declared event) were those related to the variables: Injuries Direct, Injuries Indirect, Deaths Direct, Damage Crops and Damage Property.
The rest: Income, Population, Deaths Indirect don’t show signs of being significant for the future model.

Variable	P-value	Odds-Ratio	Lower $CI_{95\%}$	Upper $CI_{95\%}$
Income	0.5322	1	1	1
Population	0.17	1	1	1
Event type(1)-Hail	0.071	1.962	0.943	4.079
Event type(2)-Other	0	4.710	2.664	8.328
Event type(3)-Flash Flood	0	9.480	5.193	17.306
Event type(4) -T.Wind	0	3.255	1.800	5.884
Event type(5)-Winter Storm	0	9.978	5.313	18.741
Direct Injuries	0.062	1.044	0.998	1.093
Indirect Injuries	0.002	1.744	1.224	2.487
Direct Deaths	0.0061	1.617	1.147	2.281
Indirect Deaths	0.0979	2.426	0.849	6.929
Damage crops	0.004	1	1	1
Damage property	0	1	1	1

Table 1: summary of bivariate analysis for each independent variable

Above the unadjusted odds-ratios are shown. It is important to highlight that the values for the continuous variables could be wrong, as the OR represents a difference in the odds for every single observation number regarding to the previous one. The value is just a reference number, but may not mean anything. For this reason, it is sometimes helpful to study the odds-ratio categorizing each continuous variable and calculating a value for every group, although choosing the ranges has to be carefully done.

When building the logistic model, one can use forward selection, adding at each step the variables considered significant by using the results from the previous bivariate analysis; or backward elimination, which is the chosen method.

There are two ways in which the latter method was conducted. Both eliminating the same variables at each step but with different finishing points. The results of the most important parameters are displayed below:

The variables eliminated in order are: Deaths Indirect, Injuries Direct, Deaths Direct, Income, Population and Damage Crops.

Model	Log-likelihood	Cox and Snells R^2	Nagelkerke R^2	Specificity	Sensitivity	Overall Percentage
1	2661.769	0.047	0.111	99.9	6.7	92.6
2	2661.830	0.047	0.111	99.9	6.7	92.6
3	2663.653	0.046	0.110	99.9	6.7	92.7
4	2665.257	0.046	0.110	99.9	6.5	92.6
5	2667.5	0.046	0.109	99.9	6.5	92.6
6	2670.22	0.045	0.107	99.9	6.5	92.6
7	2679.659	0.044	0.103	99.9	5.8	92.6

Table 2: summary of important goodness of fit parameters for each model

Model	Chi-cuadrado	gl	Sig.
1	13.669	8	0.091
2	13.835	8	0.086
3	11.894	8	0.156
4	14.261	8	0.075
5	6.511	9	0.59
6	20.314	7	0.005
7	23.758	7	0.001

Table 3: summary of Hosmer-Lemeshow test for each model

If the p-value is chosen as the decision criteria to eliminate variables, the method stops at model number 7, in which the variables eliminated are: Deaths Indirect, Injuries Direct, Deaths Direct, Income, Population and Damage Crops.
Instead, if backward elimination run by the SPSS program is studied, the method stops at step 6, where the eliminated variables are all of the above besides Damage Crops.

Looking at all the parameters, the final model chosen for the study is model number 5. It is the one that with the Hosmer_Lemeshow test is not able to reject the null hypotheses of the model fitting the data, while also getting more decent parameters for the log-likelihood (the higher the better), the broader Pseudo-R^2 values (4.6-10.9% of the dependent variable is explained by the independent variables) and a higher specificity (99.9%) and sensitivity (6.7%).

Even though it is the best model obtained, it is not a good model, the Pseudo-R^2 range is very low, the specificity too high and the sensitivity too low. Not to mention that what seem like two non-statistically significant variables are left in the model (Population with a p-value=0.118 and Damage crop 0.099).

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Population	.000	.000	2.445	1	.118	1.000	1.000	1.000
	Injuries Indirect	.441	.199	4.938	1	.026	1.555	1.053	2.295
	Event Type			97.040	5	.000			
	Event Type(1)	.671	.374	3.222	1	.073	1.956	.940	4.068
	Event Type(2)	1.386	.292	22.476	1	.000	4.000	2.255	7.095
	Event Type(3)	2.181	.309	49.850	1	.000	8.857	4.834	16.228
	Event Type(4)	1.166	.303	14.821	1	.000	3.210	1.773	5.812
	Event Type(5)	2.208	.324	46.524	1	.000	9.100	4.824	17.163
	Dge_crop	.000	.000	2.714	1	.099	1.000	1.000	1.000
	Dge_prop	.000	.000	35.210	1	.000	1.000	1.000	1.000
	Constante	-3.864	.282	187.423	1	.000	.021		

a. Variables especificadas en el paso 1: Population, Injuries Indirect , Event Type, Dge_crop, Dge_prop.

Tabla e: Final model chosen (nº5) using binary logistic regression

In the above table, it should be highlighted that between the bivariate analysis of the variables and the final binary logistic model, the values of the odds-ratios remain constant. These similarities give signs of a strong relationships of the variables in the model.

Going back to the bad results of the goodness of fit, the high specificity number stands out as a high number. This means that the negative cases (not declarations) are predicted in a very high percentage, too much compared to the positive cases (low sensitivity)

In statistics it is better to make errors type II than type I. In our model, it seems a better option to predict that a disaster will lead to a Disaster Declaration when it is not going to be the case than the other way around. Instead of trying out different thresholds, let's look at a very useful tool called the ROC curve.

The ROC curve looks at the trade-off between specifity and sensibility. All the values of the areas under the curve are around 0.5, which means that the model is not able to separate the positive group and the negative one.

This analysis confirms that our model predicts that many events will not be led to a Disaster Declaration when in reality they will be. This imbalance makes it impossible for a model to distinguish correctly among the different cases, making a weaker prediction when it seemed a decent prediction at first.

Conclusions

Logistic models are a very useful tool when predicting a dichotomous variable. The interpretation of the logistic model is very especial, as the effect in probability of each independent variable over the dependent variable is not constant. As the logarithms interfere in the prediction, the interpretation is made through the odds-ratio, the likelihood of an event taking place affected by an independent variable.

The model with the highest predictive power seemed the Model number 5, chosen even if it meant leaving two non-statistically significant variables according to the p-value criteria. These variables are Population (with a p-value=0.118) and Damage crop (p-value=0.099).

Sometimes the best option can be to leave some non-statistically significant variables in the model that just add information rather than loosing predictive power with other estimates analysed in the goodness of fit (log-likelihood, Pseudo-R2's, Hosmer-Lemeshow test, specificity and sensibility). In this case the difference was so small, that the risk could be taken.

Although the studied parameters seemed all acceptable, the specificity parameter of 99.9% showed the failure of the model to show a good number of predictions for the true positive cases, this being the predictions of the declared events, which is the main reason that this study was conducted. The sensibility of 6.5% along with the study of the ROC curves shed light over the fact that the sample size did not have enough declared events (415 against 4925, ~8%). This was the source of the problem, not enough declared events in the original data.

A possible solution to the problem could be using some of the machine learning methods as training a decision tree, oversampling or undersampling.

MEMORIA

Content

Index of Figures

Index of Tables

I. Summary

In face of a natural disaster in the United States, both county, state and federal governments are responsible for aiding citizens financially and physically to recover from the damages. The Stafford Disaster Relief and Emergency Assistance Act and its later amendments was designed to bring a systematic approach to coordinate who and when needs to provide natural disaster assistance. When a natural disaster occurs, the government of the county and state in which the natural disaster takes place must execute the state's emergency plan. On top of county and state assistance, the state government can request the US government to make a "Presidential disaster declaration", which then triggers federal financial and physical assistance through the FEMA (Federal Emergency Management Agency). This declaration is made only when certain circumstances are met (impact of disaster, scale of disaster, to name a few).

Whether a natural disaster triggers a Presidential Disaster Declaration has a big economic and financial impact on both the State government and its citizens, and therefore it is very important to understand what are the conditions or characteristics that will lead to such declaration.

The objective of this study was to build a scientific model to predict the likelihood that a given natural disaster would lead to a Presidential Disaster Declaration in the state of Maryland. The data set used in this project includes the Disaster Declarations of the past 15 years (2003-2018) in the state of Maryland by county, as well as several variables such as income, population, size of the disaster among others.

A logistic model was built using the backward elimination method, which means that all selected variables were included in the beginning model to then eliminate them one by one to understand the effect in the model. The p-value was used to define which variables were non-statistically significant, and therefore, which ones should be eliminated at each iteration. After all the iterations, there were seven different models.

The model with the highest predictive power was a model that included the following variables: population, indirect injuries, type of event, damage crops and damage property. This model has two variables that are non-statistically significant according to the p-value criteria, but the subsequent study of the goodness of fit proofed that the model was not affected by keeping these two variables. Therefore, given the parameters analysed, the two non-statistically significant variables were kept in the model.

Although the studied parameters seemed all acceptable, the specificity parameter of 99.9% showed the inability of the model to show a good number of predictions for the true positive cases (in other words, the predictions of the declared events). The sensibility of 6.5% along with the study of the ROC curves shed light over the fact that the sample size did not have enough declared events (415 against 4925, ~8%).

The problem found during this project is called imbalanced data. In other words, a specific event is very rare, happens too little as a percentage of the total number of observations. In this case, the number of declared events should have been much larger to be representative. The imbalanced data issue is very common in medical results for rare illnesses. A good method to overcome the issue of imbalanced data is by using some tools that the machine learning techniques provide.

II. Introduction

a. Context: The Presidential Disaster Declaration

Every year in the United States, hazards such as hail, thunderstorms, marine thunderstorms, heavy rain, flood, tornados, hurricanes or strong winds occur. Others of greater scale such as hurricanes or tsunamis can also happen from time to time. Whenever one of these hazards occur, the region where it took place executes a State's emergency plan to mitigate the damages and aid its citizens. (FEMA18)

The Stafford Disaster Relief and Emergency Assistance Act (Stafford Act) is a United States federal law signed in 1988 that sets a systematic and order approach on who and when should provide financial and physical assistance to aid citizens in the event of a natural disaster. (FEMA18)

Once a disaster occurs, the Governor executes the state's emergency plan. If after he believes that the state cannot cover the cost or does not have enough resources, the state government can request the US government to make a "Presidential disaster declaration", which then triggers federal financial and physical assistance through the FEMA (Federal Emergency Management Agency). (FEMA18)

Whether a natural disaster triggers a disaster declaration has a big economic and financial impact on both the State government and its citizens, as the federal government can cover up to 75% of the costs of the mitigation measures implemented[2]. This measure makes a great impact on the state's economy. (FEMA18)

The federal assistance covers a range of very different activities: coordinate all disaster relief assistance, provide the help of many collaborative entities such as the red cross and Federal agencies, assist with the distribution of food, medicines or other vital supplies, provide technical and advisory assistance to affected areas or accelerated federal assistance even if not yet requested.

b. Limitations on the current Presidential disaster declarations approach

Presidential disaster declarations in the USA are very dependent on the judgement of the President who is the person ultimately in charge of making the decision on whether to take action and help a county or a state when a hazard occurs. Furthermore, there are other factors that influence, such as how and when the Governor requests and completes the form and the reasons he/she states to request such help. (FEMA18)

One very important element in the Stafford Act described in Title III is the non-discrimination policy when providing disaster assistance.[1] As of now there is no evidence to determine if the non-discrimination clause is being enforced because there is not a scientific approach for when to make a Presidential Disaster Declaration. (FEMA18)

In sum, the current approach to determine if the President should make a Presidential Disaster Declaration is biased by subjective factors.[2] A scientific model based on different variables and past Disaster Declarations could be put in place to ensure objectivity and avoid discrimination (based on political preferences, economic power and so on). This model could define which factors and circumstances should be in place to make a Disaster Declaration. The model would quantify the different factors, and both the state and federal government could make more informed decisions. For example, the Governor would know when to request such Declaration, and the federal government could react faster.

[1] "Robert T. Stafford Disaster Relief and Emergency Assistance Act", FEMA, https://www.fema.gov/robert-t-stafford-disaster-relief-and-emergency-assistance-act-public-law-93-288-amended

[2] Stafford Disaster Relief and Emergency Assistance Act, Section 404. Hazard Mitigation (42 U.S.C. 5170c)21

III. Objective of this study

The objective of this study is to build a scientific model to predict the likelihood that a given natural disaster would lead to a Presidential Disaster Declaration in the state of Maryland. The data set used in this project includes the Disaster Declarations of the past 15 years (2003-2018) in the state of Maryland by county, as well as several variables such as income, population, damaged crops, number of deaths and injuries, size of the disaster among others.

More specifically, this project covers the following questions:

- Understand which characteristics of the data set are the best predictors of whether a Presidential Disaster Declaration will be made
- Study correlations between how much money is awarded during the disasters and the income of the specific county
- Build the model that best explains the variability in the data and therefore has the best predictive power

This model would improve the existing approach of when to make a Presidential Disaster Declaration, a very relevant and important topic given the high occurrence of natural disasters in the US and the large amounts of money at stake.

The main limitation of this study is the narrow scope of the data set (only for the state of Maryland). The original idea was to use the dataset of natural disaster for the of the US territory. However, during the process of organizing the data, an error was found in the code that transformed some of the columns. As this code was of public domain, the data had to be adjusted manually. Consequently, given that it was not possible to fix the code, the sample collection had to be reduced to the events happening only in the State of Maryland. This region was chosen considering that aparently enough events had been declared over the time period and the sample size was sufficiently big to make a predictive model.

IV. Methodology

a. Statistic model selection

The first step is to define which model to use to build the regression model: linear, logistic, or machine learning model. In this chapter we will cover the limitations of each model to then select the one that fits best for the purpose of this project.

Linear Regression

Linear Regression is a statistical method that makes relationships between independent variables and a variable that will be explained by those variables. It uses the Ordinary Least Squares method (OLS) when fitting the line, which implies finding out the line that goes through all the points with the lowest error possible. (AGRE02)

The relationship is given by the following function:

$$Y = \beta_0 + \beta_1 * X_1 + \cdots + \beta_k * X_k + u$$

Being:

- Y: dependent variable
- X_k: independent variables
- u: random disturbances
- β_i: estimated parameters

Given that in this case the dependent variable needs to answer a yes/ no question (Would a given natural disaster trigger a Presidential Disaster Declaration?), the model will have some problems that need to be considered. In a linear regression model with a binary dependent variable, the estimated Y expresses the probability of the event taking place. The problems will be the following(MART17):

1. Heteroscedasticity of the random disturbances as the constant value of the disturbances cannot be assured any more, as they depend on the value of the X.
2. The random disturbances can only take two values for each individual, so it is not possible to assure the hypotheses of the normality of the perturbances.
3. Finally, it will be possible to get probabilities greater than 1 and lower than 0, something that is not mathematically possible.

Given the limitations of the linear regression model, as the dependent variable is a binary decision, a logistic model will be considered. (AGRE02)

Logistic Regression

Within the logistic model, there are two options that can be considered: LOGIT and PROBIT. They both use a cumulative probability function in order to keep the dependent variable inside the range of [0,1]. The chosen model for this research is LOGIT, although both models in the computer give the same results. (AGRE02)

LOGIT is mathematically noted as:

$$Log[P(Y=1)] = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 * X_1 + \cdots + \beta_n * X_n$$

There are some differences that need to be taken into account:

1. The β does not measure the marginal effect any more, the importance relies on the sign that it gets, to interpret the impact (positive or negative).
2. The marginal effect is not constant any more. The slope of the function changes depending on the point. The slope represents the probability change regarding the initial probability of the variable.
3. The linear regression model was estimated using the OLS approach, but instead, the LOGIT is estimated using the Maximum Likelihood Estimation (MLE).
4. The goodness of fit cannot be studied with the R^2. One of the alternatives to be used are the Pseudo R^2 of Mac-Fadden, percentage of cases correctly predicted, sensitivity analysis or significative contrasts that will be studied at a later stage.

As mentioned earlier, this model uses the Maximum Likelihood method, which means that as the β value can not be interpreted any more, so another tool must be used. The concept used is called the Odds Ratio, a statistic that measures the effect of how likely an outcome is going to happen submitted to a certain exposure relative to the same outcome without that exposure. It's a coefficient between two odds.

$$Odds\ Ratio = \frac{\frac{P_1(Y=1)}{1-P_1(Y=1)}}{\frac{P_0(Y=1)}{1-P_0(Y=1)}} = \frac{e^{\beta_1} * e^{\beta_0+\beta_1 * X_1+\beta_2 * X_2+\cdots+\beta_n * X_n}}{e^{\beta_0+\beta_1 * X_1+\beta_2 * X_2+\cdots+\beta_n * X_n}} = e^{\beta_1}$$

The equation above shows the case for β_1, but it works for any variable i that wants to be explained. In order to interpret the number, the deduction is that:

- OR>1: the exposure leads to a greater outcome oddity
- OR<1: the exposure leads to a lower outcome oddity
- OR=1 or close: exposure does not affect outcome oddity

Besides the interpretation of parameters, it is very important to know how to choose which variables will be then considered meaningful and added to the model. Find out which ones will be the best estimators. For this reason, there are several methods of building the model to consider when the variables are being chosen: (VIDH19)

- **Forward selection**: begins with an empty model (only the $ß_0$). Then starts adding one by one the variable that either has the highest scores in some tests as for example Chi-square test or the lowest p-value. Once a variable enters the model, it remains. The model is finally built when a new added variable no longer helps it to improve.
- **Backward elimination**: in this case, all the independent variables are introduced at the beginning. Then deleted one by one based on different possible criteria, either by choosing the variable with the results given by the Wald test or removing those with the highest p-values. Once a variable exits the model, it is never put back. The model is finished when it meets the requirements considered.
- **Stepwise selection**: is a combination of the previous two. It begins with no variables. New variables are added analysing the results step by step. Unlike in the forward selection, a variable can be inserted and eliminated whenever it does not seem helpful for the model.

The guidelines on how to choose the different variables to build the logistic model are explained; also, how the results can be interpreted. But it is important to consider whether the whole sample size should be taken to fit the model. One of the problems of doing so is that there could be overfitting. (VIDH19)

Machine Learning

Machine Learning is a method of data analysis that works with huge amounts of data and can organize all the information to build a predictive model. The computer can do all this on its own, part of what is called artificial intelligence, thanks to the algorithms a human can insert onto a computer. (GUPT17)

This method will not be used for this research, although it is a very helpful tool that could be interesting. The interest relies in the methods that enable the partition of the data into randomly assigned categories that can be used for a better prediction. One of the problems of using all the available sample size was that there could be overfitting.

The way to solve this problem is by using cross-validation, a method that consist of a partition of the data into different sets: some called training data and testing data. After, a model is run with only the training data and evaluated on the testing data. Finally, the testing error is measured, and the model evaluated. (GUPT17)

Important cross-validation methods and useful for this research could be:

- **Holdout Method**: simpler way that consists of easily removing part of the sample and testing on the rest that was left out. (GUPT17)

- **K-fold**: if too much of the sample is left out, there could be a problem of underfitting. This method makes sure that enough data remains on each group (training and testing). The data is divided into k sets, one is the testing set and the other k-1 the training set al together. There are k trials where the model is estimated, so every set gets to be a testing set once, and the error is averaged k number of times. As a general set k=10. (GUPT17)
- A slight variation of this method is using the **Stratified K-fold**, in which the imbalances of the results of the data are resolved. To further understand this, if the dependent variable has a success rate of 40% in the whole bunch of data, each independent set should have that same percentage. (GUPT17)
- **Leave-p-out**: excludes p data points from the training set of data, n-p data points are used to fit the model, and the rest (p) used for testing. As the k-fold method, there will be n-p trials in order to get one error for each and then calculate the average. This method can be too extensive if p is a very large number, leading to an infeasible solution. A common approach is using Leave-One-Out method, in which p=1. (GUPT17)

After the analysis of the three potential regression models that could be used, we will use the logistic model.

b. Sample Collection

To build this model I have used three different sources of data: The Storm Events database from the National Oceanic and Atmospheric Administration (NOAA), the US Census Bureau, and a data set from University of Maryland.

- **Storm Events Database from NOAA**

The NOAA Storm Events Database compiles information about the different events (natural disasters) that have taken place in the United States since January 1950. The experts began recording data for tornadoes, but since January 1996 they began recording the data of all types of natural disasters. There are now 48 different types of events that are recorded (NWS Directive 10-1605)

The data set used for this project includes the events recorded by the NOAA for the state of Maryland from 2003 to 2017 (Annex A: Maryland_fin.xlsx).

- **US Census Bureau**

From the US Census Bureau, I extracted two data points: income and population data by county from 2013 to 2017. Every county has what is called a FIPS county code (Federal Processing Standard Publication), a unique 5-digit identification number. The first two digits refer to the state while the remaining three refer to the specific county. All these data points were added to the NOAA Storm database using the FIPS code as the reference to consolidate both data sets.

- **University of Maryland:**

The University of Maryland had a database with all the events that had been recorded as "Presidential Disaster Declaration" in the United States. By matching this dataset with the database from NOAA, I could categorize all the events as 1 (if Disaster Declaration was made) or 0 (if a Disaster Declaration was not made).

c. Data Analysis

After collecting all the data, the excel file that includes all the incidents in Maryland contains 12,386 rows. Many rows relate to the same event, and since this would biased the model, the number of rows must be reduced. (BURS08)

Using Matlab, a condition is set to do so. It consists of three logical conditions that have to happen at the same time:

- the beginning date of the event and the beginning date of the previous event need to have a time difference of less than six days
- the event type names have to match
- whether the event was declared a disaster or not has to match too (0 or 1).

It is reasonable to think that in order for an event to be considered the same, there must be a continuation of at least five days. If a disaster takes place, stops for more than five days, and then is back, it will be considered a different event. The final compilation reduces the number of rows to 5,340. Below is the code used in MatLab to compile the rows by unique events: (BURS08)

```
if ((BeginDateMatlab(i)-BeginDateMatlab(i-1))<6) && (EVENT_TYPE(i)==EVENT_TYPE(i-1) && (Disaster(i)==Disaster(i-1)))
```

Figure 1: code in Matlab to compile events showing three conditions

d. Variables description

All the variables considered for the future model will be analysed in this section:

- Income and Population (US Census Bureau)
- Direct injuries, indirect injuries, direct deaths, indirect deaths, event type, damage crops, and damage property (NOAA database)
- Presidential Disaster Declaration (University of Maryland)

Variables can be either be continuous or categorical. A description of the significance of each variable is listed below:

Income
Average household income by county (in USD). Household incomes includes all wages, salaries or any kind of transfer payments coming from the Government, such as retirement income. When the code is run to reduce the number of rows in the excel, the variable income coming from different counties is the weighted average using population size (Matlab code included in Annex B)

Population
Number of inhabitants by county. The variable population in the model expresses the people that live in the county affected by the hazard. Therefore, when some of the incidents are compiled, the cumulative population is calculated.

Direct Injuries
Continuous variable that represents the number of injuries that are directly caused by the event.

Indirect Injuries
Continuous variable that represents of the number of injuries that are indirectly caused by the event.

Direct Deaths
Continuous variable that represents the number of deaths that are directly caused by the weather event.

Indirect Deaths
Continuous variable that represents the number of deaths that are indirectly caused by the event.

Event type
As explained in the description of the Storm Events Database, there are 48 different type of events defined by the NOAA. The five most frequent event types are chosen as a variable by order: Thunderstorm Wind, Winter Weather, Flash Flood, Hail and Winter Storm. The remaining event types are gathered into the same variable called Other.

A categorical variable like this one has to be turned into a (0/1) by using dummy variables come into place. One of the categories is chosen as the base category (e.g. Thunderstorm Wind), and dichotomous variables are added until there are no more left. As a result, we now have five variables rather than six, as we have used one as our base. In the model the selection of that base category will be chosen using a specified criterion. This was just an example of how the categorization and the creation of the dummy variables would be done. (Annex B shows the MatLab code use to create the categorization).

Damage crops
Continuous variable that represents the total damage in USD done by the event to the crops.

Damage property
Continuous variable that represents the total damage in USD done by the event to the properties.

Presidential Disaster Declaration
This is the dependent variable, the one that will be predicted. It consists of a binary decision that shows whether the event has been declared a disaster or not. It shows a 1 if it has been declared and a 0 otherwise.

e. Variable Analysis

Before building the model, it is necessary to analyse the effect of each independent variable over the dependent one (Declaration). Several tests are conducted to try to prove the relationship between the variables: contingency tables, chi-tests, t-student and univariate logistic regression. (BURS08)

The categorical variables are studied using contingency tables. The tables show the frequency of every category indicating the occurrence of the outcome (yes declared or not declared), followed by a chi-squared test that states as the null hypotheses: H_0-there is no significant difference between the dependent and independent variable. In this analysis, event is the only categorical variable.

However, for the continuous variables, histograms and box plots will be studied instead, followed by a t-test for mean comparisons and a univariate logistic regression. The t-test for mean comparisons states the following hypotheses:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 != 0$$

If the null hypotheses are rejected (p-value lower than 0.05), then it will mean that both variables in the study do not have the same mean for the different possible outcomes (yes declared/not declared) and so the value of the independent variable might be greater or lower (depending on which mean is higher) in order for an event to be declared a disaster.

Then the univariate logistic regression is modelled in order to have an estimate of what is called an odds-ratio. In logistic regression, as a result of being a logarithmic function, the parameter of study betha is not constant as for linear regression. To study the constant effect of an independent variable, the odds-ratio must be studied OR=(exp(betha)), which represents the odds of being declared a disaster in each case's exposure compared to the case of not having that exposure. Important to highlight difference in odds, not probability. (BURS08)

The ultimate goal of this method is observing the possible risk factors of the model fit, a prior study that serves as a potential influence of each variable in the model.

Event type

As the event type is a categorical variable, frequency tables for each of the cases are built. In order to get a better view of the differences, all the data is gathered into a contingency table.

		Frequency	Percentage
Event type	Thunderstorm	1283	26.05%
	Winter Weather	654	13.28%
	Flash Flood	398	8.08%
	Hail	436	8.85%
	Winter Storm	242	4.91%
	Other	1912	38.82%
Total		4925	1

Table 1: frequency table of variable Event Type for not declared events

		Frequency	Percentage
Event type	Thunderstorm	83	20.00%
	Winter Weather	13	3.13%
	Flash Flood	75	18.07%
	Hail	17	4.10%
	Winter Storm	48	11.57%
	Other	179	43.13%
Total		415	1

Table 2: frequency table of variable Event Type for declared events

The frequency tables compared show evidence of some of the events tending more towards one of each of the outcomes. These are the Winter Weather and Hail, being twice as frequent of not being declared, and Winter Storm and Flash Flood, double as frequent of being declared a disaster.

		Declaration		
		Not Declared	Declared	Total
Event Type	Thunderstorm	1283 26.05%	83 20.00%	1366 25.58%
	Winter Weather	654 13.28%	13 3.13%	667 12.49%
	Flash Flood	398 8.08%	75 18.07%	473 8.86%
	Hail	436 8.85%	17 4.10%	453 8.48%
	Winter Storm	242 4.91%	48 11.57%	290 5.43%
	Other	1912 38.82%	179 43.13%	2091 39.16%
Total		4925 100.00%	415 100.00%	5340 100.00%

Table 3: contingency table for variable Event Type

Everything gathered together does not show clear evidence of any of the cases being less influential. Although, apparently, the Winter Weather is the variable that could affect less, as it is the variable with less difference in percentage towards not being declared and further from being declared. For this reason, the variable Winter Weather is the one taken

as the base category when building the dummy variables, it will be the one that will not have a constant predicted in the model and will serve as a reference group.
Only slight tendencies are shown in the frequencies towards being declared a disaster, these are three variables Winter Storm, Flash Flood and Other.

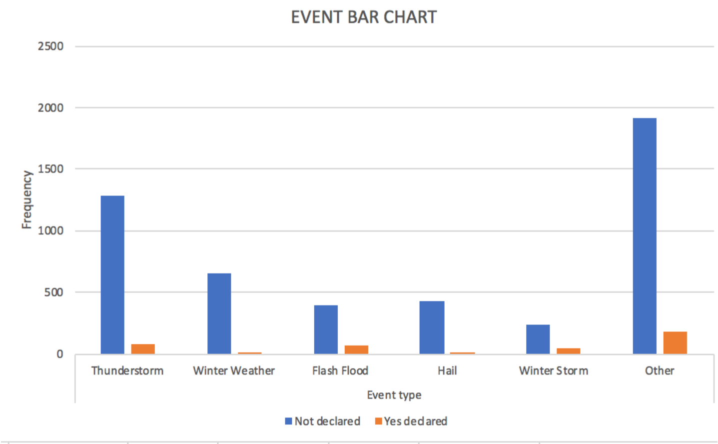


Figure 2: Even Type bar chart, frequency of two possible outcomes

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	123.375[a]	5	.000
Razón de verosimilitud	122.300	5	.000
N de casos válidos	5340		

a. 0 casillas (0.0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 22.54.

Table 4: Chi-squared test for variable Event Type

A chi-squared test result of 123.375 for a 5 degrees of freedom study shows that there is association between the variable event type and the fact of being declared a disaster. The number X^2=123.375>11.07 with a p-value of 0.05 allows the null hypotheses to be rejected (H_0: there is no significant difference between the dependent and independent variable). So, there is significant difference.

How much each category affects the overall effect is explained below by studying the value of the odds-ratio.

As the variable Winter Weather has been chosen as the base category, the constant betha won't be estimated and consequently the odds-ratio won't be calculated either. Instead, the logistic regression is set leaving the category Winter Weather as the reference group, obtaining the same results in both Matlab and SPSS. The only difference of using each program is the step of creating the dummy variables (useful for the following study of continuous variables too). The program SPSS does them on its own, but due to the number of variables that there are in this research and to simplify the method, the code is run in

Matlab to categorize the dummy variables and enabling me, the user to choose the variable I would like as the reference category. The code is presented in Annex B.

```
Estimated Coefficients:
                   Estimate       SE        tStat       pValue  
                   ________    _______    _______    __________

    (Intercept)     -3.9182    0.28009    -13.989    1.8254e-44
    x1               1.5496     0.2908      5.329     9.877e-08
    x2                 1.18    0.30213     3.9058    9.3913e-05
    x3               2.2492    0.30708     7.3245    2.3986e-13
    x4              0.67373    0.37359     1.8034      0.071326
    x5               2.3004    0.32159     7.1534     8.468e-13
```

Table 5: logistic regression for variable Event Type using Matlab

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Event Type			106.930	5	.000			
	Event Type(1)	.674	.374	3.252	1	.071	1.962	.943	4.079
	Event Type(2)	1.550	.291	28.398	1	.000	4.710	2.664	8.328
	Event Type(3)	2.249	.307	53.648	1	.000	9.480	5.193	17.306
	Event Type(4)	1.180	.302	15.255	1	.000	3.255	1.800	5.884
	Event Type(5)	2.300	.322	51.171	1	.000	9.978	5.313	18.741
	Constante	-3.918	.280	195.686	1	.000	.020		

a. Variables especificadas en el paso 1: Event Type.

Table 6: logistic regression for variable Event Type using SPSS

Both analyses show the same result, but the SPSS allows the collection of broader information in an easier way. This is the reason why, from now on, SPSS will be used in order to build the logistic models and run statistical tests. From the table, the results importance lies in the exp(Betha) column. It expresses the odds-ratio compared to the base category (Winter Weather).

The odds-ratio is greater than 1 in all of the cases by order: Hail, Other, Flash Flood, Thunderstorm Wind and Winter Storm. With a value of 1.962, 4.710, 9.48, 3.255 and 9.978 respectively, it would mean that the odds of being declared a disaster in each case is that amount of times higher than being declared in the case of a Winter Weather scenario. The higher the odds-ratio the riskier the variable is to tend the event to be declared. These strongest variables would be Flash Flood and Winter Storm. Although, it should be noted that for variable Hail the confidence interval contains the 1 in the confidence interval, so the previous statements cannot be supported.

All p-values are lower than 0.05 except for the event Hail. Once again that would mean all of the results are significant but the one given for this event, that cannot be concluded to be associated to a higher probability of being declared compared to the Winter Weather event.

Income

The income is a continuous variable. Firstly, a simple histogram is plotted in both situations, the data that leads to a disaster being declared (Figure 3) and to the opposite result (Figure 4):

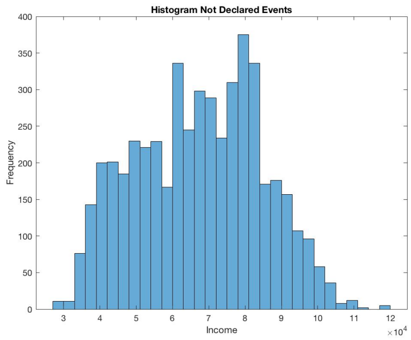


Figure 3:histogram for variable Income, not declared events

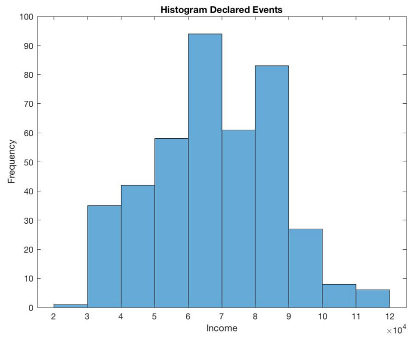


Figure 4:histogram for variable Income, declared events

The histograms, backed up by the box plot, show a normal distribution.

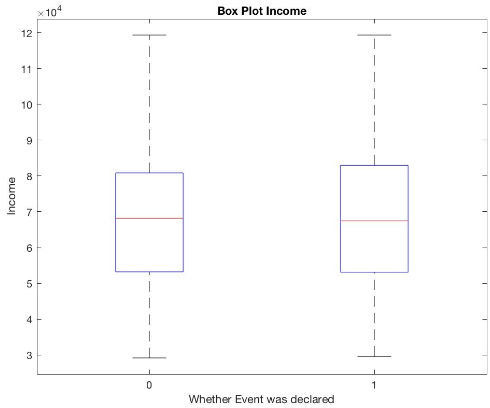


Figure 5: box-plot of variable Income for each possible outcome

	Declaration			Estadístico	Error estándar
Income	0	Media		67346,4428	247,537645
		95% de intervalo de confianza para la media	Límite inferior	66861,1587	
			Límite superior	67831,7270	
		Media recortada al 5%		67261,2698	
		Mediana		68197,2451	
		Varianza		301778811	
		Desviación estándar		17371,7820	
		Mínimo		29162,0000	
		Máximo		119386,000	
		Rango		90224,0000	
		Rango intercuartil		27642,6031	
		Asimetría		-.027	.035
		Curtosis		-.790	.070
	1	Media		67904,3038	916,201191
		95% de intervalo de confianza para la media	Límite inferior	66103,3174	
			Límite superior	69705,2902	
		Media recortada al 5%		67558,0513	
		Mediana		67466,0656	
		Varianza		348361218	
		Desviación estándar		18664,4373	
		Mínimo		29546,0000	
		Máximo		119386,000	
		Rango		89840,0000	
		Rango intercuartil		29902,6203	
		Asimetría		.174	.120
		Curtosis		-.397	.239

Table 7: output given by the SPSS with variable Income descriptive statistics

Figure 9 shows the output of the descriptive statistics of the variable Income. From now on, to avoid too many figures, the output for the rest of the variables will not be shown, just the mean numbers obtained.

The difference in the average is 557,861 (67904,3038-67346,4428) almost imperceptible in comparison to a county's income. Such small difference, along with a similar $IC_{95\%}$ suggest that there might not be significant differences between the variables Declaration and Income. To demonstrate this, a t-test for the difference in means is conducted:

Estadísticas de grupo

	Declaration	N	Media	Desviación estándar	Media de error estándar
Income	0	4925	67346,4428	17371,7820	247,537645
	1	415	67904,3038	18664,4373	916,201191

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias					95% de intervalo de confianza de la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	Inferior	Superior
Income	Se asumen varianzas iguales	1.756	.185	-.625	5338	.532	-557,86100	893,247919	-2308,9918	1193,26981
	No se asumen varianzas iguales			-.588	476.433	.557	-557,86100	949,051900	-2422,7059	1306,98391

Table 8: independent t-student mean comparison test for variable Income

The independent samples t-test states as the null hypotheses that the difference in means is zero. In the results given by the figure 10 the p-value is 0.532>0.05 which tells that the H_0 cannot be rejected, meaning that both variables in the study (Income and Declaration) might have a difference of 0 and that a county's income is not necessarily greater in order for an event to be declared a disaster. As it was being predicted with the descriptive statistics.

To further study the relationship between the variables Income and Declaration, a logistic model is built by only adding the variable Income. The results would be the following:

							95% C.I for exp(Betha)	
	Estimate/Betha	SE	Wald	gl	sig	Exp(Betha)	Inferior	Superior
Intercept	-2.597399791	0.205053719	160.451	1	9.02E-37	0.07446696		
Income	1.83E-06	2.93E-06	3.90E-01	1	0.53224687	1.00000183	1	1

Table 9: univariate logistic regression with variable Income

The results given by the logistic model built with only one continuous variable could include a great error, as the OR represents a difference in the odds for every single income number regarding to the previous one. The value is just a reference for the future model, as it gives kind of a mean of all the odd-ratios put together.

For this reason, it is useful to convert the continuous variable into a dichotomous variable. The odds-ratio gives the probability of the category in question related to that one in the base category. There is not an easy way to divide the data into different categories. For this reason, five different categories with the exact same length were chosen.

Annex G shows the whole code run in Matlab to be able to categorize de variable.

```
min_inc=min(Income);
max_inc=max(Income);
range_inc=(max_inc-min_inc)/5;
rg1_inc=min_inc;
rg2_inc=rg1_inc+range_inc;
rg3_inc=rg2_inc+range_inc;
rg4_inc=rg3_inc+range_inc;
rg5_inc=rg4_inc+range_inc;
rg6_inc=max_inc;
```

Figure 6: code range income data into five equal categories

It consists of a very simple way to divide the ranges: five categories of the same length. The method of stratification is a great challenge in statistics, as the information can be lost during the process. As we are only computing a variable analysis and not building a model, this is a risk that can be taken. Categorization is only done to have a reference of the effects of the variable. Keeping the down effects in mind, the process can be continued.

Once the categories are classified, the histograms were the following:

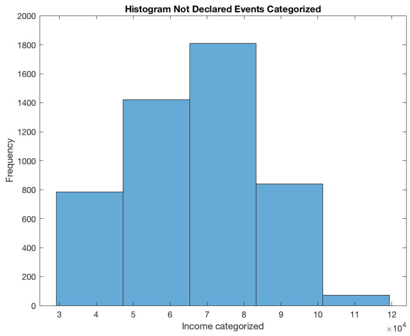


Figure 7: histogram for variable Income categorized, not declared events

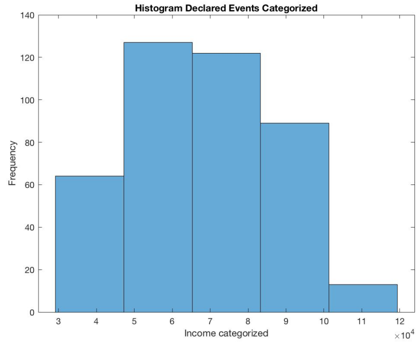


Figure 8: histogram for variable Income categorized, declared events

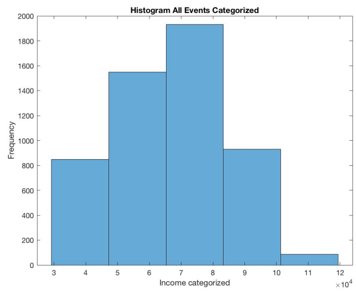


Figure 9: histogram for variable Income categorized, all events

Again, the problem is which category can be chosen as a reference group. In this case, the nominal group chosen is the category that seems to have the less tendency towards being declared a disaster. Doing this by looking at the frequencies, the chosen category is the third income range.

And the logistic results run in SPSS:

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Inc2	-.096	.151	.400	1	.527	.909	.676	1.222
	Inc3	-.377	.152	6.182	1	.013	.686	.509	.923
	Inc4	.075	.163	.208	1	.648	1.077	.782	1.484
	Inc5	-18.884	4736.787	.000	1	.997	.000	.000	.
	Constante	-2.319	.119	377.119	1	.000	.098		

a. Variables especificadas en el paso 1: Inc2, Inc3, Inc4, Inc5.

Table 10: logistic regression with variable income categorized

None of the categories seem to be significant, since the $IC_{95\%}$ for the odd-ratios contain the number 1 among them. In conclusion, the variable income does not seem to be associated at all with the fact of the event being declared a disaster.

Population

The difference in means for the variable population is 792980.79-702909.67=90071.12, being greater for the category of those events not being declared a disaster. This difference might seem big at first. Let's move onto a t-test study to check the null hypotheses.

Estadísticas de grupo

	Declaration	N	Media	Desviación estándar	Media de error estándar
Population	1	415	702909.67	939933.380	46139.515
	0	4925	792980.79	1310563,09	18674.751

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
									95% de intervalo de confianza de la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	Inferior	Superior
Population	Se asumen varianzas iguales	6.446	.011	-1.371	5338	.171	-90071.118	65715.044	-218899,45	38757.212
	No se asumen varianzas iguales			-1.810	559.490	.071	-90071.118	49775.508	-187840,82	7698.585

Table 11: independent t-student mean comparison test for variable Population

The p-value of 0.171 implies that the null hypotheses cannot be rejected, making us uncapable of rejecting the H_0 and leading to think that there might not be significant differences in the population for each possible outcome (being declared/not being declared disaster). Besides, the $IC_{95\%}$ contains the number zero.

							95% C.I for exp(Betha)	
	Estimate/Betha	SE	Wald	gl	sig	Exp(Betha)	Inferior	Superior
Intercept	-2.426856572	0.060591426	1604.227	1	0	0.088		
Population	-6.30E-08	4.59E-08	1.88E+00	1	1.70E-01	1	1	1

Table 12: univariate logistic regression with variable Population

After running the logistic model, the odds-ratios obtained give a value of 1, again this number could be mistaken, but a first impression suggests that there is no association between the variable Population and Declaration.

Once again, the variable is divided into different categories. This time, if five categories of equal sizes are chosen, the data is all gathered on the first categories, as seen in the Box Plot in figure 10, having too many samples and leaving the rest of the categories almost or completely empty.

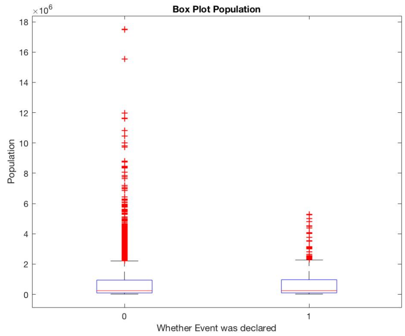


Figure 10: box plot for variable Population for each possible outcome

Taking the distribution into account, the ranges to categorize the variable were chosen using the quantiles, distributing the same size of the sample in five different categories. Again, this is not the perfect way to do it, but there is not one way that can be great of doing so.

The results of the logistic regression using the categories of population and choosing the first range as the base category were the following:

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Pop2	.172	.164	1.100	1	.294	1.188	.861	1.639
	Pop3	.102	.166	.380	1	.538	1.108	.800	1.533
	Pop4	.216	.163	1.758	1	.185	1.241	.902	1.707
	Pop5	.108	.166	.426	1	.514	1.114	.805	1.543
	Constante	-2.595	.120	470.071	1	.000	.075		

a. Variables especificadas en el paso 1: Pop2, Pop3, Pop4, Pop5.

Table 13: logistic regression with variable Population categorized

All of the $IC_{95\%}$ contain the number 1, which means that there are no evidences of any of the categories influencing in the fact of being declared a disaster. Although there is no sign of association, it is still good to add the variable to the whole logistic model as we we'll see later on.

Injuries Direct

A difference in means of (0.53-0.05) 0.48 could suggest that the variable injuries indirect has a significant impact in the variable disaster declaration. Running the t-test below:

Prueba T

Estadísticas de grupo

	Declaration	N	Media	Desviación estándar	Media de error estándar
Injures Direct	1	415	.53	7.900	.388
	0	4925	.05	1.008	.014

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
									95% de intervalo de confianza de la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	Inferior	Superior
Injures Direct	Se asumen varianzas iguales	58.976	.000	3.897	5338	.000	.479	.123	.238	.720
	No se asumen varianzas iguales			1.234	415.136	.218	.479	.388	-.284	1.242

Table 14: independent t-student mean comparison test for variable Injuries Direct

The null hypotheses of equal means are rejected, a p-value of 0 backed up by a confidence interval that does not contain 0 states that the means will be different with a 95% of confidence.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Injures Direct	.043	.023	3.480	1	.062	1.044	.998	1.093
	Constante	-2.480	.051	2339.310	1	.000	.084		

a. Variables especificadas en el paso 1: Injures Direct.

Table 15: univariate logistic regression with variable Injuries Direct

Once again, the variable studied together in the logistic model gives an odds-ratio of 1.044.

In order to simplify this case, and because many of the values have 0 injuries, only two categories are built. The first one is those events with no injuries and the other one with one or more. The graph bars are represented below:

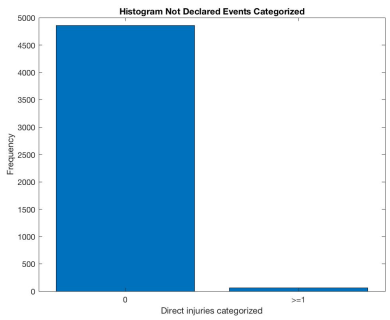


Figure 11: histogram variable Direct Injuries not declared events categorized

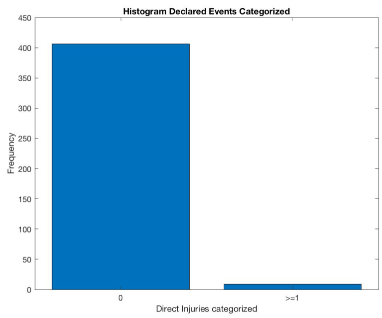


Figure 12: histogram variable Direct Injuries declared events categorized

There is a slight greater proportion of cases with one or more direct injuries in the declared events group and a greater proportion of no injuries in the not declared group.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Injdir	.521	.360	2.098	1	.148	1.684	.832	3.408
	Constante	-2.483	.052	2309.500	1	.000	.084		

a. Variables especificadas en el paso 1: Injdir.

Table 16: logistic regression with variable Injuries Direct categorized

The logistic model using the dichotomous variable Injuries Direct gives an odds-ratio of 1.684 with a $CI_{95\%}$ containing the number 1, which is not a significative OR.

Injuries Indirect

The difference in means is (0.05-0) 0.05, almost imperceptible.

Prueba T

Estadísticas de grupo

	Declaration	N	Media	Desviación estándar	Media de error estándar
Injuries Indirect	1	415	.05	.550	.027
	0	4925	.00	.101	.001

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	95% de intervalo de confianza de la diferencia Inferior	Superior
Injuries Indirect	Se asumen varianzas iguales	96.273	.000	4.900	5338	.000	.045	.009	.027	.063
	No se asumen varianzas iguales			1.679	416.347	.094	.045	.027	-.008	.098

Table 17: independent t-student mean comparison test for variable Injuries Indirect

The t-test results show that there might be a significant difference between the variables the fact of being declared a disaster and the value of the number of indirect injuries, as the p-value is equal to 0 and the confidence interval does not contain the 0.

	Estimate/Betha	SE	Wald	gl	sig	Exp(Betha)	95% C.I for exp(Betha) Inferior	Superior
Intercept	-2.481870129	0.05130713	2339.927	1	0	0.084		
Indirect Injuries	5.56E-01	1.81E-01	9.46E+00	1	0.00210527	1.744	1.224	2.487

Table 18: univariate logistic regression with variable Injuries Indirect

An OR of 1.744 is obtained for the simple logistic regression without categorizing the continuous variable.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Injind	1.922	.629	9.345	1	.002	6.838	1.993	23.454
	Constante	-2.482	.051	2336.741	1	.000	.084		

a. Variables especificadas en el paso 1: Injind.

Table 19: univariate logistic regression with variable Injuries Indirect categorized

The OR obtained when the variable is categorized gives a number of 6.838, with a confidence interval too broad. The number would mean that the odds of being declared a disaster declaration when there is more than 1injury is that times more than when there are no injuries.

Deaths Direct

The difference in means is (0.04-0.01) 0.03, at first look seems too small.

Prueba T

Estadísticas de grupo

	Declaration	N	Media	Desviación estándar	Media de error estándar
Deaths Direct	1	415	.04	.327	.016
	0	4925	.01	.162	.002

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas F	Sig.	prueba t para la igualdad de medias t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	95% de intervalo de confianza de la diferencia Inferior	Superior
Deaths Direct	Se asumen varianzas iguales	36.813	.000	3.079	5338	.002	.028	.009	.010	.046
	No se asumen varianzas iguales			1.749	431.271	.081	.028	.016	-.004	.060

Table 20: independent t-student mean comparison test for variable Deaths Direct

The t-test results reject the null hypotheses with a p-value of 0.02, leading to the conclusion that there is a significant difference in the effect of the variable direct deaths in the variable declaration.

	Estimate/Betha	SE	Wald	gl	sig	Exp(Betha)	95% C.I for exp(Betha) Inferior	Superior
Intercept	-2.484903529	0.05146328	2331.437	1	0	0.083		
Direct Deaths	4.81E-01	1.75E-01	7.51E+00	1	0.00613424	1.617	1.147	2.281

Table 21: univariate logistic regression with variable Deaths Direct

The odds-ratio given by the logistic model with the continuous variable Direct Deaths gives a number of 1.617.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B)	
								Inferior	Superior
Paso 1[a]	Deathsdir	1.079	.356	9.160	1	.002	2.941	1.463	5.915
	Constante	-2.490	.052	2318.449	1	.000	.083		

a. Variables especificadas en el paso 1: Deathsdir.

Table 22: univariate logistic regression with variable Deaths Direct categorized

When the model is run categorizing the variable, the OR jumps to 2.941. It would mean that having one or more direct deaths as explained in the definition at the beginning would increase the odds of the event happening being declared a disaster 2.941 times compared to not having any deaths.

Deaths Indirect

There is a total difference in means of (0.01-0) 0.01. Small number that does not seem to make the difference significant.

Prueba T

Estadísticas de grupo

	Declaration	N	Media	Desviación estándar	Media de error estándar
Deaths Indirect	1	415	.01	.110	.005
	0	4925	.00	.051	.001

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	95% de intervalo de confianza de la diferencia: Inferior	Superior
Deaths Indirect	Se asumen varianzas iguales	13.228	.000	1.822	5338	.069	.005	.003	.000	.011
	No se asumen varianzas iguales			.994	429.428	.321	.005	.005	-.005	.016

Table 23: independent t-student mean comparison test for variable Deaths Indirect

The t-test shows no evidence that there is a significant difference in the means, not being able to demonstrate a strong association between the variables Indirect Deaths and Declaration.

							95% C.I for exp(Betha)	
	Estimate/Betha	SE	Wald	gl	sig	Exp(Betha)	Inferior	Superior
Intercept	-2.47722992	0.05122092	2339.039	1	0	0.084		
Indirect Deaths	8.86E-01	5.35E-01	2.74E+00	1	0.097917	2.426	0.849	6.929

Table 24: univariate logistic regression with variable Deaths Indirect

The logistic model results using the continuous variable give an OR of 2.426.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Deathsindir	1.783	1.226	2.115	1	.146	5.946	.538	65.707
	Constante	-2.476	.051	2340.813	1	.000	.084		

a. Variables especificadas en el paso 1: Deathsindir.

Table 25: univariate logistic regression with variable Deaths Indirect categorized

The odds-ratio resulted from the categorized variable, contains the number 1 in the confidence interval, so the value is not relevant.

Damage Crops

The difference in means between the two different outcomes is (11951.81-61.24=11890.57) a prety big number that might suggest a strong dependence of the value of the damage of the crops in whether the event is declared a disaster or not.

Estadísticas de grupo

	Declaration	N	Media	Desviación estándar	Media de error estándar
Dge_crop	0	4925	61.24	585.047	8.337
	1	415	11951.81	132241.170	6491.464

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas F	Sig.	prueba t para la igualdad de medias t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	95% de intervalo de confianza de la diferencia Inferior	Superior
Dge_crop	Se asumen varianzas iguales	157.666	.000	-6.316	5338	.000	-11890.567	1882.659	-15581.347	-8199.787
	No se asumen varianzas iguales			-1.832	414.001	.068	-11890.567	6491.469	-24650.916	869.783

Table 26: independent t-student mean comparison test for variable Damage Crops

The t-test results validate the previous assumptions. A p-value of 0 rejects the null hypotheses of the means in the two groups being the same.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Dge_crop	.000	.000	8.175	1	.004	1.000	1.000	1.000
	Constante	-2.498	.052	2335.806	1	.000	.082		

a. Variables especificadas en el paso 1: Dge_crop.

Table 27: univariate logistic regression with variable Damage Crops

The logistic regression OR for the continuous variable Damage Crops used is 1.

For this situation, the value of 20,000 dollars for the damage was found to be a good limit value in order to transform the variable into a dichotomous one.

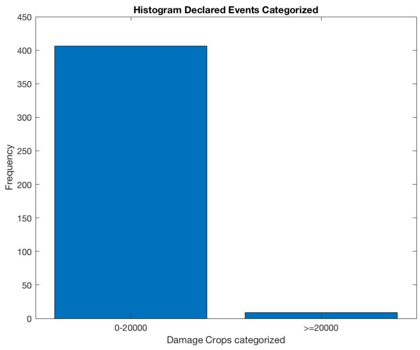


Figure 13: histogram variable Damage Crops declared events categorized

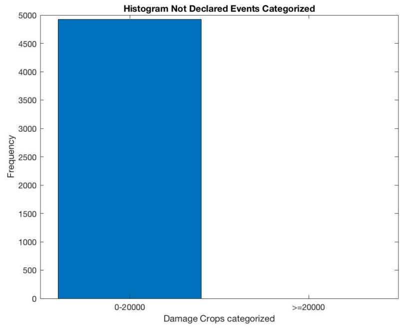


Figure 14: histogram variable Damage Crops not declared events categorized

Both histograms represent a clear situation in which all the events with a greater damage in crops of 20,000 dollars are declared a disaster.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Dmg_crops	23.699	13397.656	.000	1	.999	1,960E+10	.000	.
	Constante	-2.496	.052	2336.240	1	.000	.082		

a. Variables especificadas en el paso 1: Dmg_crops.

Table 28: univariate logistic regression with variable Damage Crops categorized

Damage Property

There is a difference in means of (1675272.1-14104.33=1661167.77). The number seems a very big and clear difference.

Estadísticas de grupo

	Declaration	N	Media	Desviación estándar	Media de error estándar
Dge_prop	0	4925	14104.33	139773.082	1991.684
	1	415	1675272.10	20638320,8	1013095,32

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	95% de intervalo de confianza de la diferencia Inferior	Superior
Dge_prop	Se asumen varianzas iguales	108.634	.000	-5.653	5338	.000	-1661167,8	293864.507	-2237262,2	-1085073,3
	No se asumen varianzas iguales			-1.640	414.003	.102	-1661167,8	1013097,28	-3652623,8	330288.264

Table 29: independent t-student mean comparison test for variable Damage Property

Looking at the t-test the results show a significant difference between the values of the variable damage property and the outcome of declaration of the event.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Dge_prop	.000	.000	47.748	1	.000	1.000	1.000	1.000
	Constante	-2.564	.053	2311.667	1	.000	.077		

a. Variables especificadas en el paso 1: Dge_prop.

Table 30: univariate logistic regression with variable Damage Property

The OR gives a value of exactly 1.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Dmg_prop	1.223	.134	83.617	1	.000	3.398	2.614	4.417
	Constante	-2.630	.057	2128.981	1	.000	.072		

a. Variables especificadas en el paso 1: Dmg_prop.

Table 31: univariate logistic regression with variable Damage Crops categorized

When analysed the dichotomous variable, the OR jumps to a value of 3.398 with a p-value of 0, which would be significant.

The results from all the different tests done for each independent variable aren't at all real. These estimates are made to have an idea of whether they are important for the future model or not. All the variables combined can give different results that will need to be compared. The logistic model will be studied afterwards.

A summary of all the unadjusted odd-ratios is shown below:

Variable	**P-value**	**Odds-Ratio**	**Lower $CI_{95\%}$**	**Upper $CI_{95\%}$**
Income	0.5322	1	1	1
Population	0.17	1	1	1
Event type(1)-Hail	0.071	1.962	0.943	4.079
Event type(2)-Other	0	4.710	2.664	8.328
Event type(3)-Flash Flood	0	9.480	5.193	17.306
Event type(4) -T.Wind	0	3.255	1.800	5.884
Event type(5)-Winter Storm	0	9.978	5.313	18.741
Direct Injuries	0.062	1.044	0.998	1.093
Indirect Injuries	0.002	1.744	1.224	2.487
Direct Deaths	0.0061	1.617	1.147	2.281
Indirect Deaths	0.0979	2.426	0.849	6.929
Damage crops	0.004	1	1	1
Damage property	0	1	1	1

Table 32: summary of all the Odd-ratios obtained from the univariate analysis

To sum up, the variables that seem to be significant after the bivariate analysis are Event Type, Injuries Direct, Injuries Indirect, Deaths Direct, Damage Crops and Damage Property.

The remaining variables, Income, Population, and Deaths Indirect do not show signs of being statistically significant for the future model.

V. Results

a. Model fitting

There are two different ways of building the logistic model: forward selection or backward elimination. The former consists of adding the variables one at a time by using the results from the previous bivariate analysis. The latter, backward elimination, is the method that will be used in this project and consists of adding all variables at once, to the remove on e by one to see the effect in the model. (LAER18)

The first step is to build the model with all the variables. One common rule is "not to choose a variable for every ten individuals studied with the outcome that wants to be analysed" (LAER18). If there are 415 events that were declared a disaster, there should not be more than 20 variables.

Then, variables will be deleted following the p-value criteria (measure of the statistical significative measure) or the Wald test (backward elimination). (LAER18)

Although some of the continuous variables were categorized in order to study them deeply, they will be added to the model as continuous. This is done in order to avoid missing out any important information.

Looking at the different p-values of the different variables, as many models as needed will be run using SPSS until the model is considered good enough. Finally, we will look at how well each model fits the observed cases and an estimate of the Pseudo-R. The latter is a measure of how much the independent variables explain a certain amount of the dependent variable, and it is measured in a percentage range (Cox and Snell's R-Squared and Nagelkerke's R-Squared).

Some analysts recommend introducing the variables to the beginning model by selecting those that during the univariate analysis showed a p-value of no more than 0.25. This is a strict rule, as some of the variables that should not be entered according to this standard might add some useful information to the model.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Income	.000	.000	2.235	1	.135	1.000	1.000	1.000
	Population	.000	.000	3.740	1	.053	1.000	1.000	1.000
	Event Type			96.727	5	.000			
	Event Type(1)	.650	.374	3.024	1	.082	1.916	.921	3.988
	Event Type(2)	1.374	.293	22.072	1	.000	3.953	2.228	7.014
	Event Type(3)	2.153	.309	48.440	1	.000	8.611	4.696	15.790
	Event Type(4)	1.146	.303	14.293	1	.000	3.147	1.737	5.702
	Event Type(5)	2.215	.324	46.839	1	.000	9.161	4.858	17.274
	Injures Direct	-.190	.167	1.297	1	.255	.827	.597	1.147
	Injuries Indirect	.409	.210	3.789	1	.052	1.506	.997	2.274
	Deaths Direct	.330	.212	2.437	1	.118	1.392	.919	2.107
	Deaths Indirect	.181	.711	.064	1	.800	1.198	.297	4.830
	Dge_crop	.000	.000	2.892	1	.089	1.000	1.000	1.000
	Dge_prop	.000	.000	35.349	1	.000	1.000	1.000	1.000
	Constante	-4.164	.348	143.521	1	.000	.016		

a. Variables especificadas en el paso 1: Income, Population, Event Type, Injures Direct, Injuries Indirect , Deaths Direct, Deaths Indirect, Dge_crop, Dge_prop.

Table 33: Model 1, all variables included

On this first model should be highlighted how the values of the OR remain almost the same for all the variables. The OR that belong to the categories of the variable Event Type show no decrease above 10% besides the value of the category Other (Hail 1.962 to 1.916; Other 4.710 to 3.953; Flash Flood 9.480 to 8.611; Thunderstorm Wind 3.255 to 3.147; Winter Storm 9.978 to 9.161. Also, the $CI_{95\%}$ remain pretty much around the previously obtained values.

The variables that have broadened their $CI_{95\%}$ are all the direct/ indirect deaths and injuries. Their p-values also show a non-significance for the model at its current state.

As the variable Deaths Indirect shows the biggest p-value, it is eliminated from the model in step 2.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Income	.000	.000	2.246	1	.134	1.000	1.000	1.000
	Population	.000	.000	3.728	1	.054	1.000	1.000	1.000
	Event Type			96.660	5	.000			
	Event Type(1)	.650	.374	3.024	1	.082	1.916	.921	3.988
	Event Type(2)	1.375	.293	22.091	1	.000	3.955	2.229	7.018
	Event Type(3)	2.154	.309	48.481	1	.000	8.618	4.700	15.802
	Event Type(4)	1.147	.303	14.315	1	.000	3.149	1.738	5.706
	Event Type(5)	2.214	.324	46.786	1	.000	9.149	4.852	17.253
	Injures Direct	-.190	.167	1.301	1	.254	.827	.597	1.146
	Injuries Indirect	.431	.199	4.707	1	.030	1.538	1.042	2.270
	Deaths Direct	.330	.212	2.435	1	.119	1.391	.919	2.106
	Dge_crop	.000	.000	2.892	1	.089	1.000	1.000	1.000
	Dge_prop	.000	.000	35.496	1	.000	1.000	1.000	1.000
	Constante	-4.165	.348	143.613	1	.000	.016		

a. Variables especificadas en el paso 1: Income, Population, Event Type, Injures Direct, Injuries Indirect , Deaths Direct, Dge_crop, Dge_prop.

Table 34: Model 2, eliminating Deaths Indirect variable

The Odd-ratios move closer to the original values obtained in the univariate analysis. Still, Injuries Direct has a p-value that is too extreme (0.254).

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Income	.000	.000	2.257	1	.133	1.000	1.000	1.000
	Population	.000	.000	3.988	1	.046	1.000	1.000	1.000
	Event Type			96.703	5	.000			
	Event Type(1)	.654	.374	3.059	1	.080	1.923	.924	4.003
	Event Type(2)	1.375	.293	22.083	1	.000	3.953	2.228	7.014
	Event Type(3)	2.160	.309	48.783	1	.000	8.671	4.730	15.896
	Event Type(4)	1.150	.303	14.380	1	.000	3.158	1.743	5.721
	Event Type(5)	2.210	.324	46.583	1	.000	9.116	4.833	17.196
	Injuries Indirect	.430	.198	4.718	1	.030	1.537	1.043	2.266
	Deaths Direct	.277	.205	1.825	1	.177	1.319	.883	1.970
	Dge_crop	.000	.000	2.854	1	.091	1.000	1.000	1.000
	Dge_prop	.000	.000	34.600	1	.000	1.000	1.000	1.000
	Constante	-4.167	.348	143.723	1	.000	.015		

a. Variables especificadas en el paso 1: Income, Population, Event Type, Injuries Indirect , Deaths Direct, Dge_crop, Dge_prop.

Table 35: Model 3, eliminating Deaths Indirect and Injuries Direct variable

Deaths Direct is the next variable to be chosen as the exit variable for the next step. With a p-value of 0.177

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Income	.000	.000	2.245	1	.134	1.000	1.000	1.000
	Population	.000	.000	3.775	1	.052	1.000	1.000	1.000
	Event Type			97.436	5	.000			
	Event Type(1)	.650	.374	3.025	1	.082	1.916	.921	3.988
	Event Type(2)	1.378	.292	22.206	1	.000	3.968	2.237	7.040
	Event Type(3)	2.164	.309	48.991	1	.000	8.705	4.749	15.955
	Event Type(4)	1.146	.303	14.283	1	.000	3.145	1.736	5.699
	Event Type(5)	2.211	.324	46.641	1	.000	9.127	4.838	17.215
	Injuries Indirect	.431	.199	4.717	1	.030	1.539	1.043	2.271
	Dge_crop	.000	.000	2.843	1	.092	1.000	1.000	1.000
	Dge_prop	.000	.000	35.452	1	.000	1.000	1.000	1.000
	Constante	-4.165	.348	143.598	1	.000	.016		

a. Variables especificadas en el paso 1: Income, Population, Event Type, Injuries Indirect , Dge_crop, Dge_prop.

Table 36: Model 3, eliminating Deaths Indirect, Injuries Direct and Deaths Direct variables

The model still shows variables that do not seem statistically significant. The next variable to exit the model should be Income, with a p-value of 0.134.

If we continue eliminating variables until all of the remaining ones are significant for the model, we end up with the following results after having eliminated by order: Deaths Indirect, Injuries Direct, Deaths Direct, Income, Population, Damage Crops.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Injuries Indirect	.436	.199	4.795	1	.029	1.547	1.047	2.286
	Event Type			96.119	5	.000			
	Event Type(1)	.674	.374	3.251	1	.071	1.961	.943	4.080
	Event Type(2)	1.413	.292	23.405	1	.000	4.109	2.318	7.285
	Event Type(3)	2.160	.309	48.982	1	.000	8.673	4.736	15.881
	Event Type(4)	1.144	.302	14.308	1	.000	3.140	1.736	5.681
	Event Type(5)	2.216	.324	46.848	1	.000	9.171	4.862	17.299
	Dge_prop	.000	.000	39.715	1	.000	1.000	1.000	1.000
	Constante	-3.922	.280	196.011	1	.000	.020		

a. Variables especificadas en el paso 1: Injuries Indirect , Event Type, Dge_prop.

Table 37: Model 7, , eliminating Deaths Indirect, Injuries Direct, Deaths Direct, Income, Population and Damage Crops

The p-values are all significant. Whether it is better to go through all this elimination of variables will be explained in the next chapter of this research. The collection of the important data is collected on the table below, that will also be explained later on.

Model	Log-likelihood	Cox and Snells R2	Nagelkerke R2	Specificity	Sensitivity	Overall Percentage
1	2661.769	0.047	0.111	99.9	6.7	92.6
2	2661.830	0.047	0.111	99.9	6.7	92.6
3	2663.653	0.046	0.110	99.9	6.7	92.7
4	2665.257	0.046	0.110	99.9	6.5	92.6
5	2667.5	0.046	0.109	99.9	6.5	92.6
6	2670.22	0.045	0.107	99.9	6.5	92.6
7	2679.659	0.044	0.103	99.9	5.8	92.6

Table 38: summary of important goodness of fit parameters for each step

The parameters will be explained in the next section of this report, as they are important measures of study of the goodness of fit.

Even though not all the variables are statistically significant, sometimes it can be good to take the risk and leave the variable in the model, as it can give information that makes other parameters better. A model with some non-significant variables could be preferred to another one which just keeps losing some other important parameters.

Before studying the goodness of fit, another model is studied by using the backward elimination method with the WALD statistic using the SPSS tools. The final model is the following:

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 6[a]	Event Type			96.985	5	.000			
	Event Type(1)	.673	.374	3.245	1	.072	1.960	.942	4.077
	Event Type(2)	1.402	.292	23.001	1	.000	4.062	2.291	7.203
	Event Type(3)	2.162	.309	49.072	1	.000	8.688	4.745	15.907
	Event Type(4)	1.139	.303	14.179	1	.000	3.124	1.727	5.652
	Event Type(5)	2.219	.324	46.980	1	.000	9.195	4.875	17.341
	Injuries Indirect	.437	.200	4.797	1	.029	1.548	1.047	2.289
	Dge_crop	.000	.000	2.769	1	.096	1.000	1.000	1.000
	Dge_prop	.000	.000	34.927	1	.000	1.000	1.000	1.000
	Constante	-3.922	.280	196.011	1	.000	.020		

Table 39: Final model using backward elimination by the Walt test in SPSS, Model 6

The program eliminated the variables by order: Deaths Indirect, Injuries Direct, Deaths Direct, Income and Population. Stopping the elimination at that point.

It should be pointed out that the variable Damage Crops is left in the model even though it has a p-value=0.096.

Model	Log-likelihood	Cox and Snells R^2	Nagelkerke R^2	Specificity	Sensitivity	Overall Percentage
1	2670.220	0.045	0.107	99.9	6.5	92.6

Table 40: goodness of fit parameters in backward elimination

The model obtained by the program using backward elimination matches the model number 6.

b. Goodness of fit

In order to know which model to choose we need to look at different parameters to understand how well our model fits the data (which model has the highest predicting power). The parameters to be considered are the following: (LAER18)

- The **log-likelihood parameter** refers to the function that maximizes to get optimal values for the estimated coefficients betha. The greater the number, the better the model.
- The **Pseudo-R^2s** are a measure that substitute the R^2 for linear regression. As the model is using the logarithm of probabilities, the range stays between 0 and 1. This is the reason why the Pseudo-R^2 are given in a range, the interval of an amount of independent variable that is explained by the dependent variables.
- **Specificity** and **sensibility**. Specificity accounts for the number of negative observations that have been well predicted. Sensitivity, on the other hand, refers to the number of positives that have been correctly predicted.
- **Hosmer-Lemeshow test** is a good measure of how the model can explain the observations. The test divides the observations in the categories and analyses how many of those cases are actually taking place and how many are expected. The chi-squared is the statistic used in the model. The null hypotheses states that the model fits the data. The rule to reject the null hypotheses will be the one given by the chi-squared limits.

The values for some of the above parameters are shown in the table below for each of the seven models built:

Model	Log-likelihood	Cox and Snells R^2	Nagelkerke R^2	Specificity	Sensitivity	Overall Percentage
1	2661.769	0.047	0.111	99.9	6.7	92.6
2	2661.830	0.047	0.111	99.9	6.7	92.6
3	2663.653	0.046	0.110	99.9	6.7	92.7
4	2665.257	0.046	0.110	99.9	6.5	92.6
5	2667.5	0.046	0.109	99.9	6.5	92.6
6	2670.22	0.045	0.107	99.9	6.5	92.6
7	2679.659	0.044	0.103	99.9	5.8	92.6

Table 41: logistic model fit eliminating variables by p-value criterion

The light green line shows the model using backward elimination using the WALD test. The numbers suggest that it is not a good model. The values are almost constant along all of the steps. Eliminating variables has almost no effect in any of the results.

The log-likelihood results show that the prediction improves as variables are being eliminated from the model.

The Pseudo-R2's show always pretty similar ranges. The numbers explain that only 4.5% to 10.7% of the dependent variable is explained by the independent variables, which could mean that there are some independent variables missing.

Looking at the specificity it seems like a strong model, as it predicts perfectly the negative results; but the sensibility (a true positive) suggests that the model is very weak, with a value of 5.8-6.7.

We now conduct the Hosmer-Lemeshow test for each of the seven models:

Model	Chi-squared	gl	Sig.
1	13.669	8	0.091
2	13.835	8	0.086
3	11.894	8	0.156
4	14.261	8	0.075
5	6.511	9	0.59
6	20.314	7	0.005
7	23.758	7	0.001

Table 42: Hosmer-Lemeshow test for each model built

The Hosmer-Lemeshow shows that model number 6 is enough to reject the null hypotheses. What would be of our interest is to have a p-value greater than 0.05 so that the null hypotheses is not rejected. This would mean that the predicted values should match the observed ones, and that the differences between them are assigned randomly.

After looking at the overall parameters, the best option for the model is number 5, a model that does not include the following variables: Deaths Indirect, Injuries Direct, Deaths Direct and Income. The model has many variables that seem significant given by the fact that their p-value is lower than 0.05 and that the OR is similar as the one studied for the bivariate analysis.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Population	.000	.000	2.445	1	.118	1.000	1.000	1.000
	Injuries Indirect	.441	.199	4.938	1	.026	1.555	1.053	2.295
	Event Type			97.040	5	.000			
	Event Type(1)	.671	.374	3.222	1	.073	1.956	.940	4.068
	Event Type(2)	1.386	.292	22.476	1	.000	4.000	2.255	7.095
	Event Type(3)	2.181	.309	49.850	1	.000	8.857	4.834	16.228
	Event Type(4)	1.166	.303	14.821	1	.000	3.210	1.773	5.812
	Event Type(5)	2.208	.324	46.524	1	.000	9.100	4.824	17.163
	Dge_crop	.000	.000	2.714	1	.099	1.000	1.000	1.000
	Dge_prop	.000	.000	35.210	1	.000	1.000	1.000	1.000
	Constante	-3.864	.282	187.423	1	.000	.021		

a. Variables especificadas en el paso 1: Population, Injuries Indirect , Event Type, Dge_crop, Dge_prop.

Table 43: Final model selected, model number 5, where Deaths Indirect, Injuries Direct, Deaths Direct and Income are eliminated

It seems reasonable to stop at step five, as the range of the Pseudo-R's is bigger and the Homer-Lemeshow results are more favourable without altering the rest of the studies. The problem is accepting a betha estimate that does not seem statistically significant (Damage Crops is left with p-value=0.099 and Population with a p-value=0.118), but is worth it if the other values of interest are not that affected, as it can be kept as an addition of information to the model.

The important thing to notice here is that the negative cases are predicted in a very high percentage, too high compared to the positive cases.

		True Condition	
		0	1
Predicted Condition	0	True negative	False Positive (Error type I)
	1	False negative (Error type II)	True Positive

Table 44: possible cases and error types

In statistics it is better to make errors type II than type I. In our model, it seems a better option to predict that a disaster will lead to a Disaster Declaration when it is not going to be the case than the other way around.

That is the reason why sometimes, the cut point in probability for the logistic model by defect is 0.5. Let's see how lowering this cut point to 0.4 affects the model:

Model 5	2667.5	0.046	0.109	99.9	6.5	92.6
Model 5b	2665.257	0.046	0.110	99.9	6.7	92.6

Table 45: owering threshold in logistic model number 5

The results are almost imperceptible. Instead of trying out different thresholds, let's look at a very useful tool called the ROC curve.

- ROC curve

The ROC curve looks at the trade-off between specificity and sensibility. In the table below we can see the values (graphs in Annex E) for each of the variables. (NARK18)

Variable	Area
Income	0.505
Population	0.507
Injuries Direct	0.504
Injuries Indirect	0.504

Deaths Direct	0.508
Deaths Indirect	0.502
Damage Crops	0.497
Damage Property	0.566

Table 46: summary of al the areas under the ROC curve for each independent variable

The ROC curve represents the relationship between de specificity and sensibility, so each point at the curve represents a decision threshold. The 45-degree line represents the points where the true positive rate (sensitivity) is equal to the false positive rate (specificity). (NARK18). In other words, in our model would mean that the proportion of correctly classified declared disasters would be the same as the proportion of incorrectly classified samples of not declared disasters.

The area under the curve gives a number between 0 and 1. When the area is 0.5, the model is not able to separate the positive group and the negative one. A 0.7 is considered a good number as it would mean that the model is able to separate the positive and negative class.

This analysis confirms that our model predicts that many events will not be led to a Disaster Declaration when in reality they will be. This imbalance makes it impossible for a model to distinguish correctly among the different cases, making a weaker prediction when it seemed a good prediction at first.

VI. Conclusion

The objective of this project was to build a predictive model to forecast whether a natural disaster would trigger a Presidential Disaster Declaration. Given that the model was to predict a binary variable (the event is declared or not), a logistic model needs to be used. A linear regression model could not be used as it is not possible to have probabilities greater than 1 and lower than 0.

In the Binomial logistic regression there are not many conditions that have to be met as compared to other models. These conditions are the following: the dependent variable is a dichotomous variable; independent variables can be either continuous, ordinal or categorical; the observations are independent; and finally, there are linear relationships between the continuous independent variables in the model and the logit function of the dependent variable. All these conditions were met in this project.

A binomial logistic model works best with dichotomous independent variables, as they are easier to interpret. In a logistic model the effect in probability of each independent variable over the dependent variable (betha) is not constant. As the logarithms interfere in the prediction, the interpretation is made through the odds-ratio(exp(betha)), the likelihood of an event taking place affected by an independent variable'exposure. It is important to note difference in odds and not probability.

The model was built using the backward elimination method, which means that all selected variables were included in the model to then eliminate them one by one to understand the effect in the model. The p-value was used to define which variables were non-statistically significant, and therefore, which ones should be eliminated at each iteration. After all the iterations, there were seven different models.

The model with the highest predictive power is model 5. The variables included in this model are population, indirect injuries, type of event, damage crops and damage property. This model has two variables that are non-statistically significant according to the p-value criteria. These variables are Population (with a p-value=0.118) and Damage crop (p-value=0.099). The subsequent study of the goodness of fit proofed that the model was not affected by keeping these two variables. Therefore, given the parameters analysed, the two non-statistically significant variables were kept in the model.

Although the studied parameters seemed all acceptable, the specificity parameter of 99.9% showed the inability of the model to show a good number of predictions for the true positive cases (in other words, the predictions of the declared events). The sensibility of 6.5% along with the study of the ROC curves shed light over the fact that the sample size did not have enough declared events (415 against 4925, ~8%). This was the source of the problem, not enough declared events in the original data.

The problem found during this project is called imbalanced data. It is common in some cases called rare events. In other words, a specific event is very rare, happens too little as a percentage of the total number of observations. In this case, the number of declared events should have been much larger to be representative. The imbalanced data issue is very common in medical results for rare illnesses. A good method to overcome the issue of imbalanced data is by using some tools that machine learning provides. The cross validation of the data can be done by previously treating the sample.

A potential solution to treat the sample data so that the model can make a better classification of the cases (true positives and true negatives) is done with the following techniques: (BURS 08) (gupt17)

- **Training a decision tree** is a technique commonly used nowadays. It classifies the data onto different groups based on some characteristics and costs. Makes hierarchies so that the categories force both decisions to take place. Furthermore, some costs in favour of the minority class, called the false negative prediction cost, can calculate with differences between clusters and added so that better results are obtained.
- **Oversampling**: adding repeated lines of the minority class.
- **Under sampling**: eliminating data sets of the majority class

The last two methods involve the risk of missing information, and therefore to mitigate this risk the data sets should be divided between a test group and a train group.

Another thing to consider is that one of the starting steps was to compile the data into fewer rows of events if some defined conditions were met. Maybe these conditions could be improved so that a stronger classification threshold is set. If an event as a tsunami takes place for example, it could be related to a flash flood happening on the next day and the Government will probably declare it as a unique event. Therefore, these conditions should be redefined to ensure that the events are consolidated into unique events in the most accurate way possible.

VII. Bibliography

Agresti, A. (2002). *Building and Applying Logistic Regression Models (Chapter 6)*. [online] Wiley Series in Probability and Statistics, p.Chapter 6. Available at: http://www.utdallas.edu/~pkc022000/6390/SP06/NOTES/Logistic_Regression_4.pdf [Accessed Apr. 2019].

Analytics Vidhya (2019). *7 Regression Types and Techniques in Data Science*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/ [Accessed Apr. 2019].

Boyle, T. (2019). *Methods for Dealing with Imbalanced Data*. [online] Medium. Available at: https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18 [Accessed Apr. 2019].

Bursac, Z., Gauss, C.H., Williams, D.K. and Hosmer, D.W. (2008). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, 3(1).

Canela, A. (2007). *Cómo hacer una Regresión Logística con SPSS© "paso a paso". (I)*. [online] Available at: http://www.fabis.org/html/archivos/docuweb/Regres_log_1r.pdf [Accessed 10 Jul. 2019].

Fema.gov. (2018). *Robert T. Stafford Disaster Relief and Emergency Assistance Act | FEMA.gov*. [online] Available at: https://www.fema.gov/robert-t-stafford-disaster-relief-and-emergency-assistance-act-public-law-93-288-amended [Accessed 10 Jul. 2019].

Gupta, P. (2017). *Cross-Validation in Machine Learning*. [online] Towards Data Science. Available at: https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f.

Ibm.com. (2014). *IBM Knowledge Center*. [online] Available at: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/regression/logistic_regression_methods.html [Accessed 9 Jul. 2019].

Investopedia. (2019). *Household Income Definition*. [online] Available at: https://www.investopedia.com/terms/h/household_income.asp [Accessed 9 Jul. 2019].

Laerd.com. (2018). *How to perform a Binomial Logistic Regression in SPSS Statistics | Laerd Statistics*. [online] Available at: https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php [Accessed May 2019].

Logisticregressionanalysis.com. (2014). *Logistic Regression Analysis – Welcome | LogisticRegressionAnalysis.com | Fast, easy guide to understanding, running, and interpreting multivariate logistic regression*. [online] Available at: http://logisticregressionanalysis.com [Accessed 9 Jul. 2019].

Low, B. (2018). *Super Simple Machine Learning — Multiple Linear Regression Part 1*. [online] Towards Data Science. Available at: https://towardsdatascience.com/super-simple-machine-learning-by-me-multiple-linear-regression-part-1-447800e8b624 [Accessed May 2019].

Martínez de Ibarreta Zorita, C., Álvarez Fernández, C., Budría Rodríguez, S., Curti González, T., Escobar Torres, L.S. and Borrás Palá, F. (2017). *Modelos Cuantitativos Para La Economia y La Empresa en 101 Ejemplos*. EV Services ed.

NCEI (2019a). *Storm Events Database | National Centers for Environmental Information*. [online] Noaa.gov. Available at: https://www.ncdc.noaa.gov/stormevents/

NCEI (2019b). *Storm Events Database | National Centers for Environmental Information*. [online] Noaa.gov. Available at: https://www.ncdc.noaa.gov/stormevents/details.jsp

Sarang Narkhede (2018). *Understanding AUC - ROC Curve*. [online] Medium. Available at: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 [Accessed Apr. 2019].

SaS. (2019). *Machine Learning: What it is and why it matters*. [online] Available at: https://www.sas.com/en_us/insights/analytics/machine-learning.html [Accessed May 2019].

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medical*, [online] pp.12–18. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3936971/ [Accessed 9 Jul. 2019].

United States Census Bureau. (2019)

US Census Bureau (2017). *State and County Estimates for 2003*. [online] Census.gov. Available at: https://www.census.gov/data/datasets/2003/demo/saipe/2003-state-and-county.html [Accessed 9 Jul. 2019].

VIII. Annexes

Annex A: table explaining the columns of the excel database

Title	Meaning	Values
BEGIN_TIME	Coded begin time of the event	numeric
END_TIME	Coded end time of the event	numerci
EPISODE_ID event_id	ID assigned by NOAA	numeric
state	State in the United States	name
state/region	Whether it is defined as a state or other type of region or tribal zone	name
EVENT_TYPE	Events permitted and defined	name
CZ_TYPE	Whether event took place county, zone or marine	C,Z,M
cz_name	County, zone or marine assigned to FIPS	name
WFO	Area of responsibility	3 letters
Begin Date Matlab	Transformed date to read in matlab	numeric
CZ_TIMEZONE	Time zone	EST,CST,MST
End Date Matlab	Transformed date to read in matlab	numeric
INJURIES_DIRECT	Injuries directly related	numeric
INJURIES_INDIRECT	Injuries indirectly related	numeric
DEATHS_DIRECT	Deaths directly related	numeric
DEATHS_INDIRECT	Deaths indirectly related	numeric
DAMAGE_PROPERTY_NUMERIC	Damage to property	numeric
DAMAGE_CROPS_NUMERIC	Damage to crops	numeric
SOURCE	Source report	name

MAGNITUDE	Measure of each event type	numeric
MAGNITUDE_TYPE	type	EG,ES,MS,MG
FLOOD_CAUSE	Reported/estimated cause flood	name
END_RANGE	Distance to geographical center event reference point	Numeric tenth of a mile
END_AZIMUTH	Compass direction from event reference point	16 point Compass possibilities
END_LOCATION	Center from which range calculated	name
BEGIN_LAT	Begin Latitud event ocurred	numeric
BEGIN_LON	Begin Longitude event ocurred	numeric
END_LAT	End Latitud event ocurred	numeric
END_LON	End Longitude event ocurred	numeric
EPISODE_NARRATIVE	Details of episode	narrative
EVENT_NARRATIVE	Details of event	narrative
DATA_SOURCE	Format source	PDS,CSV
fips	Coded FIPS for every county	Five digit number
year	Year of event	year
month	Month of event	Name month
Population	Population of each county	numeric
Income_Data	Income of each county	numeric
Urban_Rural_Designation	How far urban area	ordinal
Disaster	Whether event ws declared dister or not	Declared=1 Not Declared=0

Annex B: Matlab code explaining the compilation of variable Income

```
i=2;
j=1;
for i=2:12386
    if (( CONDITIONS
        j=j-1;
        %no necesario cambiar begin date
        EndDate(j)=EndDateMatlab(i);
        %pop stays just as the row before in order to do a weighted ratio
        %no cambio pop j todavia me viene de antes
        %we make weighted average of household inc
        Income(j)=(Income_Data(i)*Population(i)+Income(j)*Pop(j))/(Population(i)+Pop(j));
        %ya cambio pop para acumularla asi q pop es acumulativo
        Pop(j)=Pop(j)+Population(i);%es la acuml de personas
        Inj_dir(j)=INJURIES_DIRECT(i)+Inj_dir(j);
        Inj_ind(j)=INJURIES_INDIRECT(i)+Inj_ind(j);
        Deaths_dir(j)=DEATHS_DIRECT(i)+Deaths_dir(j);
        Deaths_ind(j)=DEATHS_INDIRECT(i)+Deaths_ind(j);
        Event_type(j)=EVENT_TYPE(i); %por defecto me quedo el i no seria necesario cambiarlo
        Declaration(j)=Disaster(i);
    else

         Income(j)=Income_Data(i);
```

Figure 15: code in Matlab explaining the compilation of the variable Income

Annex C: histograms of variable Event type before and after compiling database

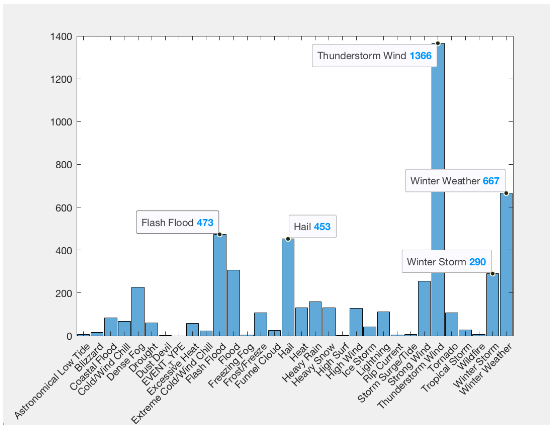


Figure 16: histogram Event Type variable before compiling

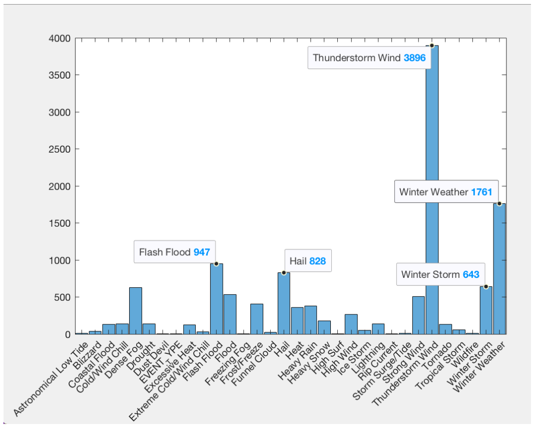


Figure 17: histogram Event Type variable after compiling

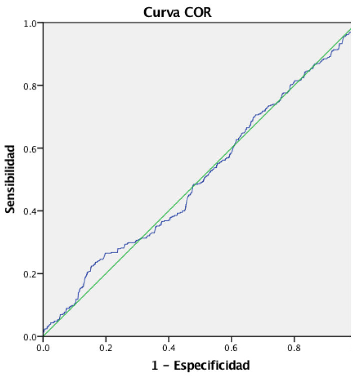
Annex D: Matlab code, how to transform a categorical variable into dummy variables

The code that made the categorization is shown in figure 18 below.

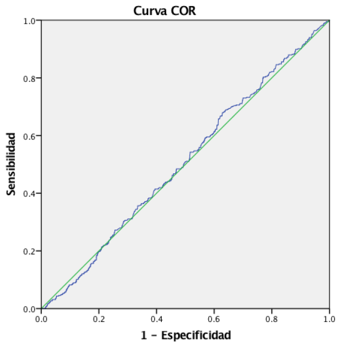
```
Events=zeros(5340,5);
i=1;
for i=1:5340
    if Event_type(i)=='Winter Weather'
        Events(i,2)=1;
    elseif Event_type(i)=='Flash Flood'
        Events(i,3)=1;
    elseif Event_type(i)=='Hail'
        Events(i,4)=1;
    elseif Event_type(i)=='Winter Storm'
        Events(i,5)=1;
    elseif Event_type(i)=='Thunderstorm Wind'
    else
        Event_type(i)='Other';
        Events(i,1)=1;
    end
    i=i+1;
end
```

Figure 18: loop code that made categorization of variable Event Type

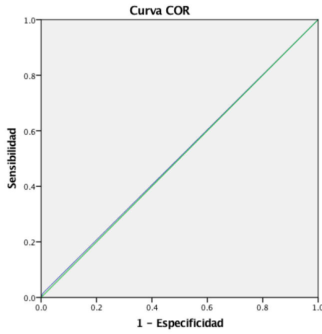
Annex E: ROC curves for each independent variable



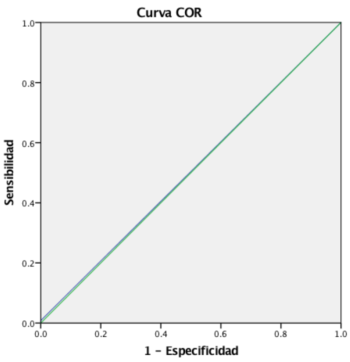
Income



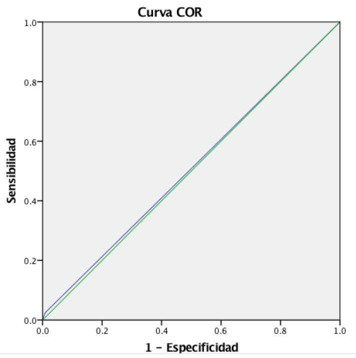
Population



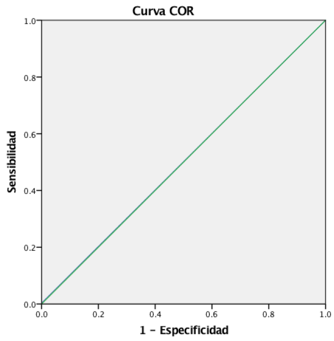
Injuries Direct



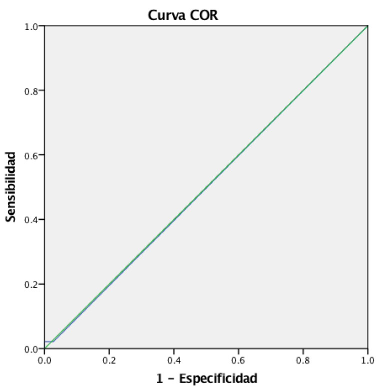
Injuries Indirect



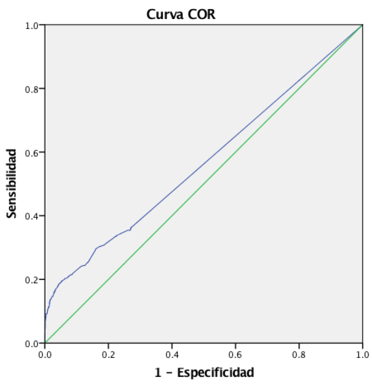
Deaths Direct



Deaths Indirect



Damage crops



Damage Property

Figure 19: ROC curves for each independent variable

Annex F: whole Matlab code, compilation of data, categorizing variable event type and logistic regression model

```
%Matlab Code_Main code
%Predictive Models for Disaster Declarations in the US
%Maria Araujo Perez
%July 2019

clear all

%LOAD SAMPLES
load Population
load Income
load fips
load Damage_Crops
load Damage_property
load Deaths_Direct
load Deaths_Indirect
load Injuries_Direct
load Injuries_Indirect
load Magnitude
load Magnitude_type
load event_type
load Urban_rural_design
load WFO
load year
load disaster
load EndDateMatlab
load BeginDateMatlab

%%
%Compile lines to gather events
%Choose a difference lower than 6 days, also event and declaration
equal

i=2;   %to compare and begin on the second position
j=1;
for i=2:12386
    if ((BeginDateMatlab(i)-BeginDateMatlab(i-1))<6) &&
(EVENT_TYPE(i)==EVENT_TYPE(i-1) && (Disaster(i)==Disaster(i-1)))
        j=j-1;
        %not necessary to change Begin Date
        EndDate(j)=EndDateMatlab(i);
        %pop stays just as the row before in order to do a weighted
ratio
        %no cambio pop j todavia me viene de antes
        %we make weighted average of household inc

Income(j)=(Income_Data(i)*Population(i)+Income(j)*Pop(j))/(Population(
i)+Pop(j));
        %ya cambio pop para acumularla asi q pop es acumulativo
        Pop(j)=Pop(j)+Population(i);%es la acuml de personas
        Inj_dir(j)=INJURIES_DIRECT(i)+Inj_dir(j);
        Inj_ind(j)=INJURIES_INDIRECT(i)+Inj_ind(j);
        Deaths_dir(j)=DEATHS_DIRECT(i)+Deaths_dir(j);
        Deaths_ind(j)=DEATHS_INDIRECT(i)+Deaths_ind(j);
        Event_type(j)=EVENT_TYPE(i); %por defecto me quedo el i no
seria necesario cambiarlo
        Declaration(j)=Disaster(i);
        Damage_crops(j)=DAMAGE_CROPS_NUMERIC(i)+Damage_crops(j);
        Damage_prop(j)=DAMAGE_PROPERTY_NUMERIC(i)+Damage_prop(j);
```

```matlab
    else
        BeginDate(j)=BeginDateMatlab(i);
        EndDate(j)=EndDateMatlab(i);
        Pop(j)=Population(i);
        Income(j)=Income_Data(i);
        Inj_dir(j)=INJURIES_DIRECT(i);
        Inj_ind(j)=INJURIES_INDIRECT(i);
        Deaths_dir(j)=DEATHS_DIRECT(i);
        Deaths_ind(j)=DEATHS_INDIRECT(i);
        Event_type(j)=EVENT_TYPE(i);
        Declaration(j)=Disaster(i);
        Damage_crops(j)=DAMAGE_CROPS_NUMERIC(i);
        Damage_prop(j)=DAMAGE_PROPERTY_NUMERIC(i);
    end
    j=j+1;
    i=i+1;
end

BeginDate=transpose(BeginDate);
EndDate=transpose(EndDate);
Pop=transpose(Pop);
Income=transpose(Income);
Inj_dir=transpose(Inj_dir);
Inj_ind=transpose(Inj_ind);
Deaths_dir=transpose(Deaths_dir);
Deaths_ind=transpose(Deaths_ind);
Event_type=transpose(Event_type);
Declaration=transpose(Declaration);
Damage_crops=transpose(Damage_crops);
Damage_prop=transpose(Damage_prop);

%%
%loop que meta en una nueva variable los nombres de los eventos con
max frecuencia
%lets pick the 5 more frequent and make the 6th other

figure(1);
histogram(EVENT_TYPE)
tbl = tabulate(EVENT_TYPE);
figure(2);
histogram(Event_type)
tbl = tabulate(Event_type);
%frequency hasn't changed


Events_before=zeros(12386,5);
i=2;
for i=2:12386
    if EVENT_TYPE(i)=='Thunderstorm Wind'
        Events_before(i,2)=1;
    elseif EVENT_TYPE(i)=='Flash Flood'
        Events_before(i,3)=1;
    elseif EVENT_TYPE(i)=='Hail'
        Events_before(i,4)=1;
    elseif EVENT_TYPE(i)=='Winter Storm'
        Events_before(i,5)=1;
    elseif EVENT_TYPE(i)=='Winter Weather'
    else
        EVENT_TYPE(i)='Other';
        Events_before(i,1)=1;
    end
```

```
    i=i+1;
end

%NO trasladar todo una fila arriba y Event_type ya trasladado
Events=zeros(5340,5);
i=1;
for i=1:5340
    if Event_type(i)=='Thunderstorm Wind'
        Events(i,2)=1;
    elseif Event_type(i)=='Flash Flood'
        Events(i,3)=1;
    elseif Event_type(i)=='Hail'
        Events(i,4)=1;
    elseif Event_type(i)=='Winter Storm'
        Events(i,5)=1;
    elseif Event_type(i)=='Winter Weather'
    else
        Event_type(i)='Other';
        Events(i,1)=1;
    end
    i=i+1;
end


%%
%LOGISTIC MODEL

%Before filtrar data

X1=[Income_Data,Population,INJURIES_DIRECT,INJURIES_INDIRECT,DEATHS_DI
RECT,DEATHS_INDIRECT,Events_before];
mdl_before =
fitglm(X1,Disaster,'Distribution','binomial','Link','logit');


%After filtrar data
X2=[Income,Pop,Inj_dir,Inj_ind,Deaths_dir,Deaths_ind,Events,Damage_cro
ps,Damage_prop,Events];
mdl= fitglm(X2,Declaration,'Distribution','binomial','Link','logit');
```

Annex G: Matlab code for the variable analysis

```
%Matlab Code_Variable Analysis
%Predictive Models for Disaster Declarations in the US
%Maria Araujo Perez
%July 2019

%%
%Event types
%tbl_Event da la contingency table de cada evento
[tbl_TS,chi2_TS,p_TS,labels_TS]=crosstab(Declaration,Events(:,1));
[tbl_WW,chi2_WW,p_WW,labels_WW]=crosstab(Declaration,Events(:,2));
[tbl_FF,chi2_FF,p_FF,labels_FF]=crosstab(Declaration,Events(:,3));
[tbl_Hail,chi2_Hail,p_Hail,labels_Hail]=crosstab(Declaration,Events(:,
4));
[tbl_WS,chi2_WS,p_WS,labels_WS]=crosstab(Declaration,Events(:,5));


%X=Events;
mdl_events =
fitglm(Events,Declaration,'Distribution','binomial','Link','logit');
%%
%Income hacer rangos

min_inc=min(Income);
max_inc=max(Income);
range_inc=(max_inc-min_inc)/5;
rg1_inc=min_inc;
rg2_inc=rg1_inc+range_inc;
rg3_inc=rg2_inc+range_inc;
rg4_inc=rg3_inc+range_inc;
rg5_inc=rg4_inc+range_inc;
rg6_inc=max_inc;

n1_inc_not=0;
n1_inc_yes=0;
n2_inc_not=0;
n2_inc_yes=0;
n3_inc_not=0;
n3_inc_yes=0;
n4_inc_not=0;
n4_inc_yes=0;
n5_inc_not=0;
n5_inc_yes=0;


j=1;
k=1;

for i=1:5340
    if ((Income(i)>=rg1_inc) && (Income(i)<=rg2_inc) &&
(Declaration(i)==0))
        n1_inc_not=n1_inc_not+1;
        inc_not(j)=Income(i);
        j=j+1;
    elseif ((Income(i)>=rg1_inc) && (Income(i)<=rg2_inc) &&
(Declaration(i)==1))
        n1_inc_yes=n1_inc_yes+1;
        inc_yes(k)=Income(i);
        k=k+1;
```

```
    elseif ((Income(i)>rg2_inc) && (Income(i)<=rg3_inc) &&
(Declaration(i)==0))
        n2_inc_not=n2_inc_not+1;
        inc_not(j)=Income(i);
        j=j+1;
    elseif ((Income(i)>rg2_inc) && (Income(i)<=rg3_inc) &&
(Declaration(i)==1))
        n2_inc_yes=n2_inc_yes+1;
        inc_yes(k)=Income(i);
        k=k+1;
    elseif ((Income(i)>rg3_inc) && (Income(i)<=rg4_inc) &&
(Declaration(i)==0))
        n3_inc_not=n3_inc_not+1;
        inc_not(j)=Income(i);
        j=j+1;
    elseif ((Income(i)>rg3_inc) && (Income(i)<=rg4_inc) &&
(Declaration(i)==1))
        n3_inc_yes=n3_inc_yes+1;
        inc_yes(k)=Income(i);
        k=k+1;
    elseif (Income(i)>rg4_inc && Income(i)<=rg5_inc &&
Declaration(i)==0)
        n4_inc_not=n4_inc_not+1;
        inc_not(j)=Income(i);
        j=j+1;
    elseif ((Income(i)>rg4_inc) && (Income(i)<=rg5_inc) &&
(Declaration(i)==1))
        n4_inc_yes=n4_inc_yes+1;
        inc_yes(k)=Income(i);
        k=k+1;
    elseif ((Income(i)>rg5_inc) && (Income(i)<=rg6_inc) &&
(Declaration(i)==0))
        n5_inc_not=n5_inc_not+1;
        inc_not(j)=Income(i);
        j=j+1;
    elseif ((Income(i)>rg5_inc) && (Income(i)<=rg6_inc) &&
(Declaration(i)==1))
        n5_inc_yes=n5_inc_yes+1;
        inc_yes(k)=Income(i);
        k=k+1;
    end
end



%aprender a poner los rangos y la frecuencia y diagrama cajas

figure()
histogram('BinEdges',[rg1_inc,rg2_inc,rg3_inc,rg4_inc,rg5_inc,rg6_inc]
,'BinCounts',[n1_inc_not,n2_inc_not,n3_inc_not,n4_inc_not,n5_inc_not])
title('Histogram Not Declared Events Categorized')
xlabel('Income categorized')
ylabel('Frequency')

figure()
histogram('BinEdges',[rg1_inc,rg2_inc,rg3_inc,rg4_inc,rg5_inc,rg6_inc]
,'BinCounts',[n1_inc_yes,n2_inc_yes,n3_inc_yes,n4_inc_yes,n5_inc_yes])
title('Histogram Declared Events Categorized')
xlabel('Income categorized')
ylabel('Frequency')
```

```
figure()
histogram('BinEdges',[rg1_inc,rg2_inc,rg3_inc,rg4_inc,rg5_inc,rg6_inc]
,'BinCounts',[n1_inc_not+n1_inc_yes,n2_inc_not+n2_inc_yes,n3_inc_not+n
3_inc_yes,n4_inc_not+n4_inc_yes,n5_inc_not+n5_inc_yes])
title('Histogram All Events Categorized')
xlabel('Income categorized')
ylabel('Frequency')

figure
boxplot(Income,Declaration)
title('Box Plot Income')
xlabel('Whether Event was declared')
ylabel('Income')


A=[n1_inc_not-n1_inc_yes,n2_inc_not-n2_inc_yes,n3_inc_not-
n3_inc_yes,n4_inc_not-n4_inc_yes,n5_inc_not-n5_inc_yes]
M_inc=max(A);
%n should be the reference group
n_inc=find(A==M_inc);

inc_cat=zeros(5340,4);
i=1;

for i=1:5340
    if ((Income(i)>=rg1_inc) && (Income(i)<=rg2_inc))
        %do nothing it's base category
    elseif ((Income(i)>rg2_inc) && (Income(i)<=rg3_inc))
        inc_cat(i,1)=1;
    elseif ((Income(i)>rg3_inc) && (Income(i)<=rg4_inc))
        inc_cat(i,2)=1;
    elseif (Income(i)>rg4_inc && Income(i)<=rg5_inc)
        inc_cat(i,3)=1;
    elseif ((Income(i)>rg5_inc) && (Income(i)<=rg6_inc))
        inc_cat(i,4)=1;
    end
end

%%
%Population
%Population hacer rangos
min_pop=min(Pop);
max_pop=max(Pop);
% range_pop=(max_pop-min_pop)/5;

rg1_pop=74585;
rg2_pop=159884.4;
rg3_pop=525304;
rg4_pop=1142382.2;


n1_pop_not=0;
n1_pop_yes=0;
n2_pop_not=0;
n2_pop_yes=0;
n3_pop_not=0;
n3_pop_yes=0;
n4_pop_not=0;
n4_pop_yes=0;
```

```
n5_pop_not=0;
n5_pop_yes=0;


j=1;
k=1;

j=1;
k=1;
for i=1:5340
    if ((Pop(i)<=rg1_pop) && (Declaration(i)==0))
        n1_pop_not=n1_pop_not+1;
        pop_not(j)=Pop(i);
        j=j+1;
    elseif ((Pop(i)<=rg1_pop) && (Declaration(i)==1))
        n1_pop_yes=n1_pop_yes+1;
        pop_yes(k)=Pop(i);
        k=k+1;
    elseif ((Pop(i)>rg1_pop) && (Pop(i)<=rg2_pop) &&
(Declaration(i)==0))
        n2_pop_not=n2_pop_not+1;
        pop_not(j)=Pop(i);
        j=j+1;
    elseif ((Pop(i)>rg1_pop) && (Pop(i)<=rg2_pop) &&
(Declaration(i)==1))
        n2_pop_yes=n2_pop_yes+1;
        pop_yes(k)=Pop(i);
        k=k+1;
    elseif ((Pop(i)>rg2_pop) && (Pop(i)<=rg3_pop) &&
(Declaration(i)==0))
        n3_pop_not=n3_pop_not+1;
        pop_not(j)=Pop(i);
        j=j+1;
    elseif ((Pop(i)>rg2_pop) && (Pop(i)<=rg3_pop) &&
(Declaration(i)==1))
        n3_pop_yes=n3_pop_yes+1;
        pop_yes(k)=Pop(i);
        k=k+1;
    elseif ((Pop(i)>rg3_pop) && (Pop(i)<=rg4_pop) &&
(Declaration(i)==0))
        n4_pop_not=n4_pop_not+1;
        pop_not(j)=Pop(i);
        j=j+1;
    elseif ((Pop(i)>rg3_pop) && (Pop(i)<=rg4_pop) &&
(Declaration(i)==1))
        n4_pop_yes=n4_pop_yes+1;
        pop_yes(k)=Pop(i);
        k=k+1;
    elseif ((Pop(i)>rg4_pop) && (Declaration(i)==0))
        n5_pop_not=n5_pop_not+1;
        pop_not(j)=Pop(i);
        j=j+1;
    elseif ((Pop(i)>rg4_pop) && (Declaration(i)==1))
        n5_pop_yes=n5_pop_yes+1;
        pop_yes(k)=Pop(i);
        k=k+1;
    end
end

B=[n1_pop_not-n1_pop_yes,n2_pop_not-n2_pop_yes,n3_pop_not-
n3_pop_yes,n4_pop_not-n4_pop_yes,n5_pop_not-n5_pop_yes]
```

```
M_pop=max(B);
%n should be the reference group
n_pop=find(B==M_pop);

pop_cat=zeros(5340,4);
i=1;
for i=1:5340
    if (Pop(i)<=rg1_pop)
        %do nothing it's base category
        pop_cat(i,:)=0;
    elseif ((Pop(i)>rg1_pop) && (Pop(i)<=rg2_pop))
        pop_cat(i,1)=1;
    elseif ((Pop(i)>rg2_pop) && (Pop(i)<=rg3_pop))
        pop_cat(i,2)=1;
    elseif ((Pop(i)>rg3_pop) && (Pop(i)<=rg4_pop))
        pop_cat(i,3)=1;
    else
        pop_cat(i,4)=1;
    end
end

%%
%Injuries Direct hacer rangos

n1_injdir_not=0;
n1_injdir_yes=0;
n2_injdir_not=0;
n2_injdir_yes=0;

j=1;
k=1;

for i=1:5340
    if ((Inj_dir(i)==0) && (Declaration(i)==0))
        n1_injdir_not=n1_injdir_not+1;
        injdir_not(j)=Inj_dir(i);
        j=j+1;
    elseif ((Inj_dir(i)==0) && (Declaration(i)==1))
        n1_injdir_yes=n1_injdir_yes+1;
        injdir_yes(k)=Inj_dir(i);
        k=k+1;
    elseif ((Inj_dir(i)>=1) && (Declaration(i)==0))
        n2_injdir_not=n2_injdir_not+1;
        injdir_not(j)=Inj_dir(i);
        j=j+1;
    elseif ((Inj_dir(i)>=1) && (Declaration(i)==1))
        n2_injdir_yes=n2_injdir_yes+1;
        injdir_yes(k)=Inj_dir(i);
        k=k+1;
    end
end


bar(c,[n1_injdir_not+n1_injdir_yes,n2_injdir_not+n2_injdir_yes])
title('Histogram All Events Categorized')
xlabel('Injuries Direct')
ylabel('Frequency')


injdir=zeros(5340,1);
```

```
i=1;
for i=1:5340
    if (Inj_dir(i)==0)
        %do nothing it's base category
        injdir(i,1)=0;
    elseif Inj_dir(i)>=1
        injdir(i,1)=1;
    end
end

%%
%%%
%Injuries Indirect hacer rangos

n1_injindir_not=0;
n1_injindir_yes=0;
n2_injindir_not=0;
n2_injindir_yes=0;

j=1;
k=1;

for i=1:5340
    if ((Inj_ind(i)==0) && (Declaration(i)==0))
        n1_injindir_not=n1_injindir_not+1;
        injindir_not(j)=Inj_ind(i);
        j=j+1;
    elseif ((Inj_ind(i)==0) && (Declaration(i)==1))
        n1_injindir_yes=n1_injindir_yes+1;
        injindir_yes(k)=Inj_ind(i);
        k=k+1;
    elseif ((Inj_ind(i)>=1) && (Declaration(i)==0))
        n2_injindir_not=n2_injindir_not+1;
        injindir_not(j)=Inj_ind(i);
        j=j+1;
    elseif ((Inj_ind(i)>=1) && (Declaration(i)==1))
        n2_injindir_yes=n2_injindir_yes+1;
        injindir_yes(k)=Inj_ind(i);
        k=k+1;
    end
end


c=categorical({'0 or 1', '>1'})
bar(c,[n1_injindir_not,n2_injindir_not])
title('Histogram Not Declared Events Categorized')
xlabel('Direct injuries categorized')
ylabel('Frequency')

c=categorical({'0 or 1', '>1'})
bar(c,[n1_injindir_yes,n2_injindir_yes])
title('Histogram Declared Events Categorized')
xlabel('Income categorized')
ylabel('Frequency')

injindir=zeros(5340,1);
i=1;
for i=1:5340
    if (Inj_ind(i)==0)
        %do nothing it's base category
        injindir(i,1)=0;
```

```
    elseif Inj_ind(i)>=1
        injindir(i,1)=1;
    end
end

%%
%Deaths Direct and Deaths Indirect are done exactly the same

%%
%Damage crops

n1_crops_not=0;
n1_crops_yes=0;
n2_crops_not=0;
n2_crops_yes=0;

j=1;
k=1;

for i=1:5340
    if ((Damage_crops(i)<20000) && (Declaration(i)==0))
        n1_crops_not=n1_crops_not+1;
        crops_not(j)=Damage_crops(i);
        j=j+1;
    elseif ((Damage_crops(i)<20000) && (Declaration(i)==1))
        n1_crops_yes=n1_crops_yes+1;
        crops_yes(k)=Damage_crops(i);
        k=k+1;
    elseif ((Damage_crops(i)>=20000) && (Declaration(i)==0))
        n2_crops_not=n2_crops_not+1;
        crops_not(j)=Damage_crops(i);
        j=j+1;
    elseif ((Damage_crops(i)>=20000) && (Declaration(i)==1))
        n2_crops_yes=n2_crops_yes+1;
        crops_yes(k)=Damage_crops(i);
        k=k+1;
    end
end

figure()
c=categorical({'0-20000', '>=20000'})
bar(c,[n1_crops_not,n2_crops_not])
title('Histogram Not Declared Events Categorized')
xlabel('Damage Crops categorized')
ylabel('Frequency')

figure()
c=categorical({'0-20000', '>=20000'})
bar(c,[n1_crops_yes,n2_crops_yes])
title('Histogram Declared Events Categorized')
xlabel('Damage Crops categorized')
ylabel('Frequency')



bar(c,[n1_crops_not+n1_crops_yes,n2_crops_not+n2_crops_yes])
title('Histogram All Events Categorized')
xlabel('Income categorized')
ylabel('Frequency')
```

```
i=1;
for i=1:5340
    if (Damage_crops(i)<20000)
        %do nothing it's base category
        crops(i,1)=0;
    elseif Damage_crops(i)>=20000
        crops(i,1)=1;
    end
end

%%
%Property

n1_prop_not=0;
n1_prop_yes=0;
n2_prop_not=0;
n2_prop_yes=0;

j=1;
k=1;

for i=1:5340
    if ((Damage_prop(i)<20000) && (Declaration(i)==0))
        n1_prop_not=n1_prop_not+1;
        prop_not(j)=Damage_prop(i);
        j=j+1;
    elseif ((Damage_prop(i)<20000) && (Declaration(i)==1))
        n1_prop_yes=n1_prop_yes+1;
        prop_yes(k)=Damage_prop(i);
        k=k+1;
    elseif ((Damage_prop(i)>=20000) && (Declaration(i)==0))
        n2_prop_not=n2_prop_not+1;
        prop_not(j)=Damage_prop(i);
        j=j+1;
    elseif ((Damage_prop(i)>=20000) && (Declaration(i)==1))
        n2_prop_yes=n2_prop_yes+1;
        prop_yes(k)=Damage_prop(i);
        k=k+1;
    end
end


i=1;
for i=1:5340
    if (Damage_prop(i)<20000)
        %do nothing it's base category
        prop(i,1)=0;
    elseif Damage_prop(i)>=20000
        prop(i,1)=1;
    end
end
```