UNIVERSIDAD DE EL SALVADOR

FACULTAD MULTIDISCIPLINARIA DE OCCIDENTE

DEPARTAMENTO DE MATEMATICAS

REGRESIÓN LOGÍSTICA MULTINOMIAL

CARRERA:

LICENCIATURA EN ESTADÍSTICA.

ASIGNATURA:

PROYECTOS DE ESTUDIOS ESTADÍSTICOS

DOCENTE:

LICDO. JAIME ISAAC PEÑA MEJIA

PRESENTADO POR:

NAHUN ANTONIO MARTÍNEZ OLMOS (MO19005)

FECHA:

25 DE SEPTIEMBRE DE 2023

Tabla de contenidos

Introducción **3**

1. Regresión Logística Multinomial en Python **3**

1.1. Análisis exploratorio . . . 4

1.1.1. Gestión de valores nulos . . . 5

1.1.2. Análisis descriptivo . . . 7

1.2. Regresión logística multinomial con scikit-learn . . . 11

1.2.1. Estimación del modelo de elección discreta . . . 11

1.2.2. Evaluación del ajuste del modelo . . . 12

1.2.3. Interpretación de los coeficientes de regresión . . . 17

1.2.4. Predicciones . . . 19

2. Regresión Logística Multinomial en R **22**

2.1. Análisis exploratorio . . . 22

2.1.1. Gestión de valores nulos . . . 22

2.1.2. Análisis descriptivo . . . 24

2.2. Estimación del modelo de elección múltiple . . . 29

2.2.1. Contraste de hipótesis para el modelo estimado . . . 32

2.2.2. Interpretación de los coeficientes de regresión . . . 34

2.2.3. Evaluación del ajuste del modelo . . . 35

2.2.4. Capacidad discriminante del modelo . . . 36

2.2.5. Predicciones . . . 37

3. Regresión Logística Multinomial en IBM SPSS **38**

3.1. Análisis exploratorio . . . 38

3.1.1 Análisis descriptivo . . . 38

3.2. Estimación del modelo de elección múltiple . . . 42

3.2.1 Contraste de hipótesis para el modelo estimado . . . 44

3.2.2. Interpretación de los coeficientes de regresión . . . 45

3.2.3. Evaluación del ajuste del modelo . . . 45

3.2.4. Capacidad discriminante del modelo . . . 46

Introducción

La regresión logística multinomial es una generalización de la regresión logística binaria, ya que tienen la misma lógica y el mismo fin, con la diferencia de que en la regresión logística multinomial las alternativas en la variable dependiente metrica son al menos 3. Esta técnica se clasifica en las técnicas de dependencia debido a que se identifica de manera clara cuál es la variable dependiente cuando se tiene una situación donde sea aplicable, y muchas veces suele confundirse con el análisis discriminante, ya que en ambas técnicas se tienen los grupos estimados en los cuales se clasificarán las nuevas observaciones, pero la regresión logística multinomial, a diferencia del análisis discriminante no es muy rigorosa con supuestos de normalidad multivariante u otras características en donde el análisis discriminante si lo es. En esta ocasión el problema que se tiene es estimar un modelo de regresión logística multinomial con el fin de predecir el origen geográfico de automóviles, con base a su potencia, aceleración, peso y número de cilindros, y se realizarán análisis descriptivos para evaluar la relación entre la variable dependiente y las variables independientes, así como también se evaluará la capacidad predictiva del modelo, y se realizarán unas predicciones a nuevas observaciones, o en este caso a nuevos automóviles, ya que uno de los objetivos de las técnicas predictivas es predecir con base a las variables explicativas o independientes.

1. Regresión Logística Multinomial en Python

A continuación, se importarán en python, todas las librerías necesarias para realizar el análisis.

```
import seaborn as sns
import numpy as np
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn import preprocessing
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score
from sklearn.metrics import confusion_matrix
import matplotlib as mpl
import matplotlib.pyplot as plt
import statsmodels.api as sm
```

1.1. Análisis exploratorio

Siempre antes de tener la intuición de que se puede llevar a cabo la aplicación de una tecnica multivariante a una determinada situación o fenomeno es muy importante realizar analisis explorios a los datos con los que se trabajará, para que de esa manera podamos tener una fuerte base al momento de decidir si vale la pena aplicar la tecnica multivariante o no, debido a que en el análisis exploratorio y descriptivo nos daremos cuenta como están nuestros datos y si cumplen con ciertos criterios dependiendo del modelo estadístico que se decee estimar, en este caso es un modelo elección multiple el que se decea estimar pero antes de decidir hacerlo o no veremos si las variables independientes perecen tener una influencia significativa sobre la variable dependiente no metrica del modelo que se pretenderá ajustar.

Ahora se importará la base de datos que contiene las variables necesarias para realizar el ajuste del modelo de elección múltiple, también se observará cuáles son las variables que contiene la base de datos y cuantos registros tiene.

```
carros = pd.read_spss('Coches.sav') #importando la base de datos
carros.shape #dimensiones de la de la base de datos
```

```
(406, 9)
```

Como podemos observar en el resultado anterior, la base de datos contiene 406 registros y 9 variables, de las cuales se muestra su descripción a continuación.

```
carros.dtypes
```

```
consumo      float64
motor        float64
cv           float64
peso         float64
acel         float64
año          float64
origen      category
cilindr     category
derivada    category
dtype: object
```

Podemos observar, se cuenta con las variables métricas consumo, motor, potencia (cv), peso, aceleración (acel) y año, mientras que las variables no métricas con las que se cuenta son origen, número de cilindros que tienen (cilind) y si el auto fue seleccionado o no. La variable explicada será origen, las variables explicativas, métricas que se utilizarán son peso, aceleración y potencia, mientras que como factor se decidirá si se utilizará

la variable cilindr o derivada, cuya decisión se basará en los resultados descriptivos que se obtengan.

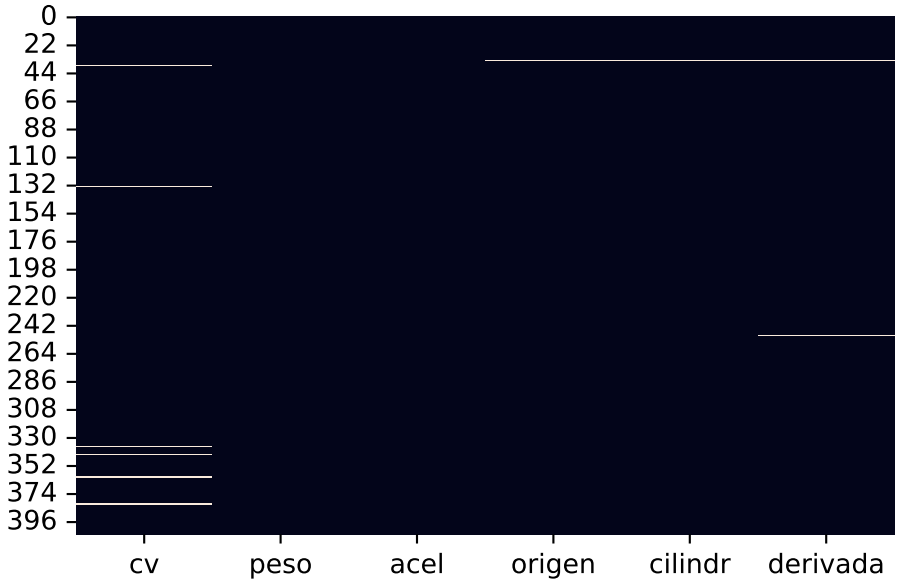
Ahora se creará un sub Data Frame con las variables que se usarán.

```
autos=carros[['cv','peso','acel','origen','cilindr','derivada']]
```

1.1.1. Gestión de valores nulos

A continuación, se mostrará de manera gráfica si las variables que estamos estudiando contienen valores nulos.

```
sns.heatmap(autos.isnull(), cbar=False)
```



Como se puede verificar al observar la figura anterior, solamente en las variables peso y aceleración no se cuenta con valores nulos, lo que significa que se imputarán los valores nulos en las restantes 4 variables. A continuación los imputaremos en las variables discretas, sustituyendo los valores perdidos por la moda.

```
autos['origen'].fillna(autos['origen'].mode()[0], inplace=True)
```

```
<string>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.
```

```
autos['cilindr'].fillna(autos['cilindr'].mode()[0], inplace=True)
```

```
<string>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.
```

```
autos['derivada'].fillna(autos['derivada'].mode()[0], inplace=True)
```

```
<string>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.
```

Ahora imputaremos los valores perdidos de la variable potencia mediante el modelo predictivo KNN.

```
from sklearn.impute import KNNImputer


# Construimos el modelo
imputer = KNNImputer(n_neighbors=5, weights='uniform')


#Ajustamos el modelo e imputamos los missing values de la variable potencia
imputer.fit(autos[['cv']])
```

```
KNNImputer()
```

```
autos['cv'] = imputer.transform(autos[['cv']]).ravel()
```

```
<string>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.
```

por último se comprobará que ya no existen valores nulos en ningunas de las variables a utilizar.

```
#Suma de valores nulos en cada variable
autos.isnull().sum()
```

```
cv          0
peso        0
acel        0
origen      0
cilindr     0
derivada    0
dtype: int64
```

Debido a que se desea ajustar un modelo de elección múltiple para explicar la relación entre el origen geográfico de los automóviles con las variables independientes o para predecir el origen de los automóviles mediante las variables independientes (potencia, número de cilindros, aceleración, peso y derivada)

1.1.2. Análisis descriptivo

Para poder determinar cuáles variables independientes valen la pena agregar a nuestro modelo de elección múltiple se analizará de manera bi variada la relación entre la variable dependiente (origen) con las variables independientes, mediante tablas de contingencia se analizará la relación entre la variable dependiente (origen) y las variables no métricas (número de cilindros y derivada) y mediante gráficos de cajas se analizará la relación entre las variables independientes métricas (peso, aceleración y potencia) y la variable dependiente origen, y estos gráficos se harán por categorías de la variable origen. A continuación empezaremos analizando la relación entre la variable origen y la variable número de cilindros.

```
data_crosstab = pd.crosstab(autos['origen'], autos['cilindr'], margins = True)
print(data_crosstab)
```

cilindr	3 cilindros	4 cilindros	5 cilindros	6 cilindros	8 cilindros	All
origen						
EE.UU.	0	73	0	74	107	254
Europa	0	66	3	4	0	73
Japón	4	69	0	6	0	79
All	4	208	3	84	107	406

Como podemos observar en la tabla anterior, parece ser que los automóviles de 4 cilindros son de los que se tienen cantidades más o menos similares en cada una de las 3 posibles regiones de origen (Estados Unidos, Europa o Japón), y que sí, un automóvil es de 8 cilindros prácticamente es seguro que su origen sea Estados unidos; sin embargo, los automóviles de Estados Unidos parece ser que se distribuyen de manera uniforme en cuanto a los que son de 4 y 6 cilindros. En conclusión, se nota que esta variable puede aportarnos mucha información para el modelo logístico que se desea estimar, por lo que esta variable será incluida como variable

explicativa. Ahora se analizará la relación entre la variable origen y la variable derivada (Filtro).

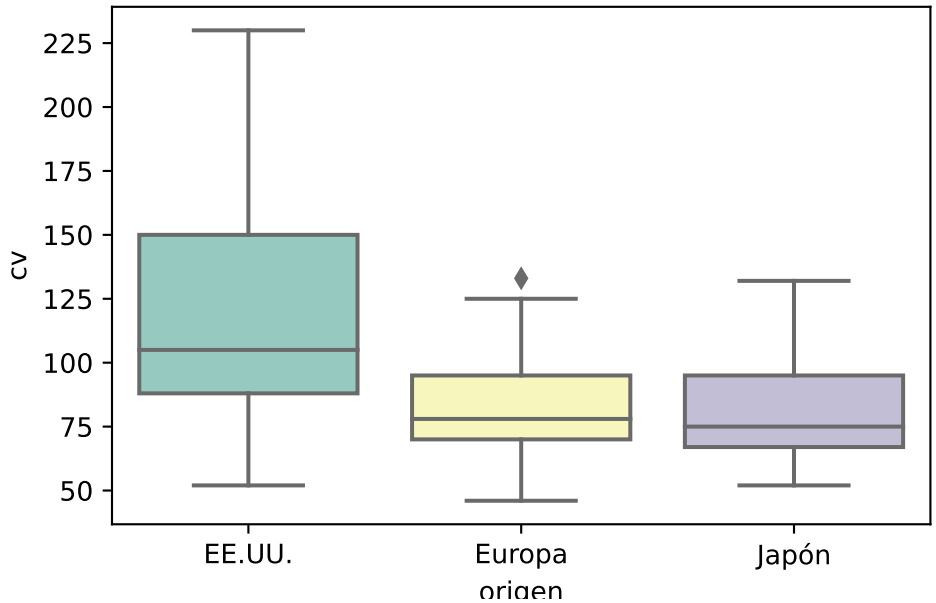
```
data_crosstab = pd.crosstab(autos['origen'], autos['derivada'], margins = True)
print(data_crosstab)
```

```
derivada  No Seleccionado  Seleccionado  All
origen
EE.UU.                107           147  254
Europa                   0            73   73
Japón                    0            79   79
All                    107           299  406
```

Como podemos observar en el resultado anterior, parece ser que los automóviles de origen Europa y Japón son seleccionados prácticamente el 100% mientras que los de origen en Estados unidos son seleccionados aproximadamente el 50% (esto puede restar potencialidad a la capacidad predictiva del modelo), es por esta razón que la variable filtro (derivada) se considera que no vale la pena incluirla en nuestro modelo.

A continuación seguiremos con el análisis de la variable potencia, ya que es una variable métrica, en la cual se analizará que tanto varían las potencias de los automóviles con respecto a los orígenes de los mismos, estas diferencias se muestran en la siguiente figura.

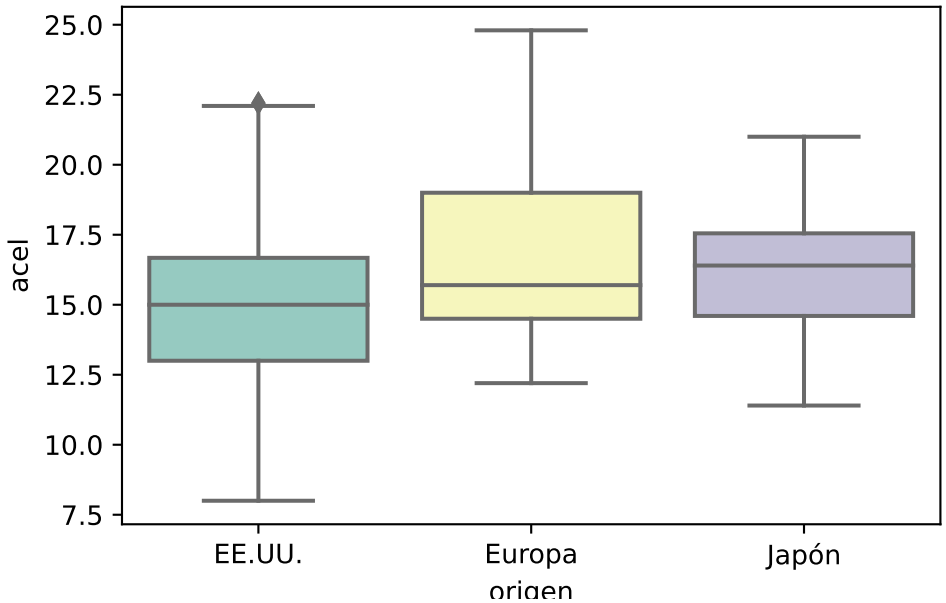
```
#import seaborn as sns


# Box plot por grupo
sns.boxplot(x = autos['origen'], y = autos['cv'], palette = "Set3")
```



Como podemos observar en la figura anterior, parece ser que las potencias medias de los vehículos con respecto a su origen varían, aunque la mayor diferencia se observa con los automóviles de Estados Unidos, los cuales difieren y varían más con respecto a los de Europa y Japón, que es poco lo que difieren y varían entre ellos, por esta razón se ha decidido tomar en cuenta esta variable como variable explicativa.

Ahora se analizará si vale la pena incluir la variable aceleración como variable explicativa en nuestro modelo de respuesta cualitativa, esto se analizará de la misma manera que con la variable potencia. A continuación se muestra el respectivo gráfico de la aceleración con respecto a su origen.

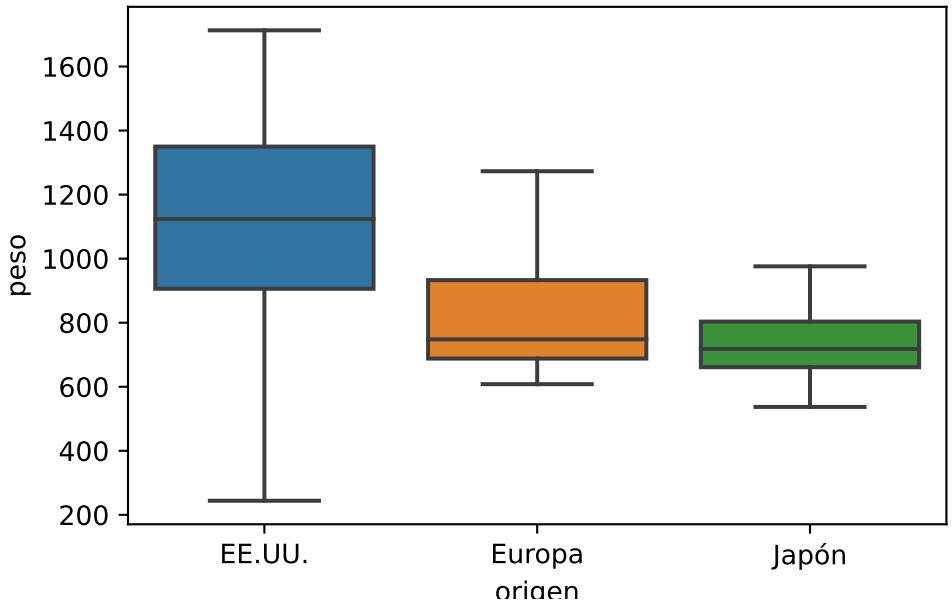
```
sns.boxplot(x = autos['origen'], y = autos['acel'], palette = "Set3")
```



Como podemos observar en la figura anterior, se observa que la diferencia en la aceleración medio de los automóviles con origen en Europa difiere aproximadamente igual con los de origen en Japón y Estados Unidos, y la diferencia entre la aceleración media es más grande entre Japón y Estados unidos que las anteriormente mencionadas, también cabe mencionar que en Europa es donde se da la aceleración promedio más alta en los automóviles, debido a estas diferencias descritas se decide por tomar en cuenta la variable aceleración para nuestro modelo de respuesta múltiple.

Ahora analizaremos la discriminación del origen con respecto al peso de los vehículos, para realizar este análisis se muestra a continuación, un boxplot del peso de los automóviles con respecto a su origen.

```
sns.boxplot(x = autos['origen'], y = autos['peso'])
```



Como podemos observar en la figura anterior, los pesos medios difieren con respecto a su origen, es decir, el origen de los automóviles tiene influencia en su peso, si miramos los gráficos de Estados Unidos y Japón ni siquiera se traslapan, por ello se decide que si vale la pena agregar la variable peso a nuestro modelo de respuesta múltiple.

1.2. Regresión logística multinomial con scikit-learn

A continuación, se pasarán las variables independientes categóricas a variables dummies (esto es esencial para el ajuste del modelo).

```
cilindro = pd.get_dummies(carros["cilindr"])
```

1.2.1. Estimación del modelo de elección discreta

A continuación se estimará el modelo de regresión logística con origen como la variable dependiente y con las variables peso, aceleración, potencia, cilindros y filtro (derivada) como independientes, y se dividirá el conjunto de datos en una parte de entrenamiento y una de prueba.

```
automoviles=pd.read_csv('automoviles.csv')


x = automoviles.drop(['origen','derivada','cilindr'], axis=1)
x = pd.concat([x,cilindro], axis=1)
```

```
#x['cilindr']=pd.Categorical(x['cilindr'])

#x['Seleccionado']=pd.Float64Dtype(x['Seleccionado'])


y = automoviles['origen']
trainX, testX, trainY, testY = train_test_split(x, y, test_size = 0.2)
```

A continuación se muestra el ajuste del modelo de regresión logística multinomial.

```
log_reg = LogisticRegression(solver='newton-cg', multi_class='multinomial')
log_reg.fit(trainX, trainY)
```

```
LogisticRegression(multi_class='multinomial', solver='newton-cg')
```

```
y_pred = log_reg.predict(testX)
```

1.2.2. Evaluación del ajuste del modelo

A continuación, se muestra la precisión y la tasa de error de nuestro modelo logístico.

```
print('Accuracy: {:.2f}'.format(accuracy_score(testY, y_pred)))
```

```
Accuracy: 0.65
```

```
print('Error rate: {:.2f}'.format(1 - accuracy_score(testY, y_pred)))
```

```
Error rate: 0.35
```

Como podemos observar en el resultado anterior, nuestro modelo tiene una precisión aproximadamente del 70%, lo que significa que es buen ajuste el que se ha realizado.

Ahora se mostrarán las puntuaciones de validación cruzada para nuestro modelo.

```
clf = LogisticRegression(solver='newton-cg', multi_class='multinomial')
scores = cross_val_score(clf, trainX, trainY, cv=5)
scores
```

```
array([0.66153846, 0.64615385, 0.75384615, 0.72307692, 0.734375  ])
```

Como podemos observar, las puntuaciones de validación cruzada son altas para cada una de las variables explicativas, lo que da aún más solides a nuestro de modelo logístico multinomial.

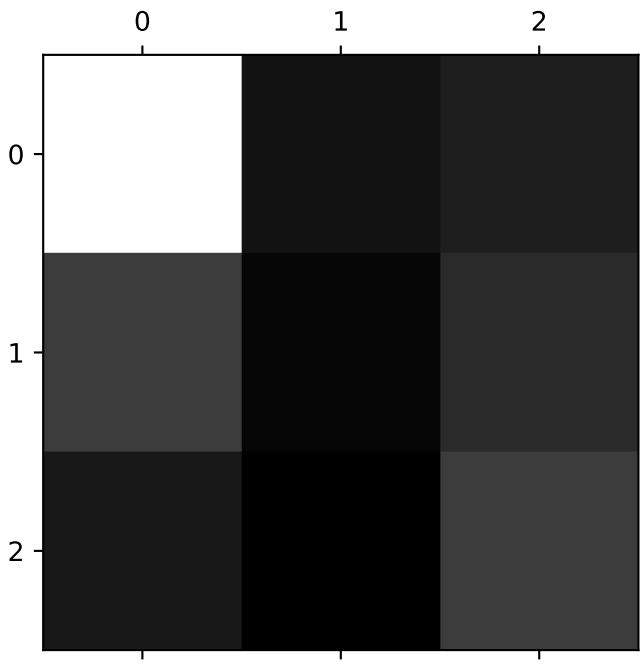
Uno de los mejores indicadores para evaluar el ajuste de un modelo logístico multinomial es la matriz de confusión, la cual nos muestra la cantidad total de predicciones acertadas de nuestro modelo logit, para ello se muestran la cantidad acertada para cada posible origen al que un automóvil puede pertenecer cuando se conoce su potencia, su peso, su aceleración, su número de cilindros y si pasó o no el filtro, por ello a continuación se muestra la matriz de confusión de nuestro modelo.

```
confusion_matrix = confusion_matrix(testY, y_pred)
print(confusion_matrix)
```

```
[[42  3  5]
 [10  1  7]
 [ 4  0 10]]
```

Como se puede observar, son muchos más los automóviles que se han clasificado de manera correcta por nuestro modelo que los que se han clasificado de manera correcta, esto muestra el poder predictivo de nuestro modelo. A continuación se mostrarán de manera gráfica la matriz los resultados de la matriz de confusión.

```
plt.matshow(confusion_matrix, cmap=plt.cm.gray)
plt.show()
```



Como se puede visualizar en la figura anterior, es para la clase 0 (Estados Unidos) que más se han acertado las predicciones realizadas por nuestro modelo, seguido por la clase 2 (Japón) y por último la clase 2 (Europa).

En este caso, el modelo de regresión logística multinomial calcula la probabilidad de que un determinado automóvil pertenezca a un determinado origen, estas probabilidades son complementarias (suman 1), y al origen que el automóvil tenga más probabilidad de pertenecer es el destino predicho donde el automóvil se clasifica ya cuando hemos ingresado al modelo de respuesta cualitativa sus características que describen su perfil, a continuación se muestran 6 de estas probabilidades.

```
probability = log_reg.predict_proba(testX)
for i in range(6):
  k=list(list(probability)[i])
  mensaje='Probabilidad de pertenecer a uno de los 3 posibles orígenes'
  o1='Estados Unidos='+str(round(k[0],4))
  o2,o3= '; Europa='+str(round(k[1],4)),'; Japón='+str(round(k[2],4))
  print(mensaje+' (automovil '+str(i+1)+'):\n'+o1+o2+o3+'\n')
```

```
Probabilidad de pertenecer a uno de los 3 posibles orígenes (automovil 1):
Estados Unidos=0.5178; Europa=0.2595; Japón=0.2228
```

```
Probabilidad de pertenecer a uno de los 3 posibles orígenes (automovil 2):
Estados Unidos=0.9961; Europa=0.0018; Japón=0.0021

Probabilidad de pertenecer a uno de los 3 posibles orígenes (automovil 3):
Estados Unidos=0.9948; Europa=0.0051; Japón=0.0001

Probabilidad de pertenecer a uno de los 3 posibles orígenes (automovil 4):
Estados Unidos=0.5657; Europa=0.1956; Japón=0.2387

Probabilidad de pertenecer a uno de los 3 posibles orígenes (automovil 5):
Estados Unidos=0.5391; Europa=0.4475; Japón=0.0134

Probabilidad de pertenecer a uno de los 3 posibles orígenes (automovil 6):
Estados Unidos=0.7904; Europa=0.081; Japón=0.1286
```

Ahora mostraremos que la longitud de predicciones que se calcularon anteriormente (aunque solo se mostraron 6) es la misma que la longitud de los datos de prueba del modelo.

```
print(probability.shape[0])
```

```
82
```

```
print(testX.shape[0])
```

```
82
```

A continuación se muestran 10 casos en donde el modelo define cual es la clase predicha con base en las probabilidades.

```
df = pd.DataFrame(log_reg.predict_proba(testX), columns=log_reg.classes_)
df['predicted_class'] = y_pred
print(df.head(10))
```

```
     EE.UU.    Europa     Japón predicted_class
0  0.517751  0.259474  0.222775          EE.UU.
1  0.996070  0.001784  0.002146          EE.UU.
2  0.994814  0.005114  0.000072          EE.UU.
3  0.565695  0.195567  0.238739          EE.UU.
```

```
4  0.539105  0.447501  0.013394          EE.UU.
5  0.790385  0.081048  0.128567          EE.UU.
6  0.319511  0.545933  0.134556          Europa
7  0.895918  0.013787  0.090295          EE.UU.
8  0.383603  0.275398  0.341000          EE.UU.
9  0.312966  0.313452  0.373581           Japón
```

A continuación se mostrarán 10 casos en los cuales se observa cuál es el origen real y cuál es el origen predicho por el modelo logístico multinomial.

```
df['actual_class'] = testY.to_frame().reset_index().drop(columns='index')
print(df.head(10))
```

```
    EE.UU.    Europa     Japón predicted_class actual_class
0  0.517751  0.259474  0.222775          EE.UU.       Europa
1  0.996070  0.001784  0.002146          EE.UU.       EE.UU.
2  0.994814  0.005114  0.000072          EE.UU.       EE.UU.
3  0.565695  0.195567  0.238739          EE.UU.       Europa
4  0.539105  0.447501  0.013394          EE.UU.       Europa
5  0.790385  0.081048  0.128567          EE.UU.       EE.UU.
6  0.319511  0.545933  0.134556          Europa       EE.UU.
7  0.895918  0.013787  0.090295          EE.UU.       Europa
8  0.383603  0.275398  0.341000          EE.UU.        Japón
9  0.312966  0.313452  0.373581           Japón       Europa
```

Como podemos observar en los 10 resultados anteriores, el modelo ha acertado en 9 de ellos.

Ahora lo que se realizará es la comprobación de la plausibilidad de los orígenes predichos, es decir, si se predijeron correctamente, si el resultado de la resta fue 0 (columna check), eso significará que el modelo ha realizado de manera correcta la predicción. A continuación se muestran los primeros 5 resultados.

```
le = preprocessing.LabelEncoder()
df['label_pred'] = le.fit_transform(df['predicted_class'])
df['label_actual'] = le.fit_transform(df['actual_class'])
df['check'] = df['label_actual'] - df['label_pred']
print(df.head())
```

```
    EE.UU.    Europa     Japón  ... label_pred label_actual  check
0  0.517751  0.259474  0.222775  ...          0            1      1
```

```
1  0.996070  0.001784  0.002146  ...          0            0      0
2  0.994814  0.005114  0.000072  ...          0            0      0
3  0.565695  0.195567  0.238739  ...          0            1      1
4  0.539105  0.447501  0.013394  ...          0            1      1

[5 rows x 8 columns]
```

A continuación, después de haber generado los valores mostrados anteriormente, los cuales indican el acierto o desacierto de las predicciones, se calculará de manera manual la precisión de nuestro modelo multinomial.

```
df['correct_prediction?'] = np.where(df['check'] == 0, 'True', 'False')
df = df.drop(['label_pred', 'label_actual', 'check'], axis=1)

true_predictions = df[(df["correct_prediction?"] == 'True')].shape[0]
false_predictions = df[(df["correct_prediction?"] == 'False')].shape[0]
total = df["correct_prediction?"].shape[0]

print('Precisión calculada manualmente:', round((true_predictions / total * 100),4))
```

```
Precisión calculada manualmente: 64.6341
```

Como podemos observar en el anterior resultado, la precisión que se ha calculado de manera manual es de aproximadamente 70% lo que significa que el modelo tiene una buena capacidad predictiva y se espera que esa capacidad se proyecte para cuando se necesite predecir el origen geográfico de un determinado automóvil.

```
wrong_pred = df[(df["correct_prediction?"] == 'False')]
```

1.2.3. Interpretación de los coeficientes de regresión

```
x = automoviles.drop(['origen','derivada','cilindr'], axis=1)
#x=pd.concat([x,filtro], axis=1)
x=pd.concat([x,cilindro], axis=1)
y = automoviles['origen']
#x=x.astype(float)

x = sm.add_constant(x, prepend = False)

mnlogit_mod = sm.MNLogit(y, x)
```

```
mnlogit_fit = mnlogit_mod.fit()
```

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.596985
         Iterations: 35

C:\Users\pc1\AppData\Local\R-MINI~1\envs\R-RETI~1\lib\site-packages\statsmodels\base\model.py:607: Conv
  warnings.warn("Maximum Likelihood optimization failed to "
```

```
print (mnlogit_fit.summary())
```

```
                          MNLogit Regression Results
==============================================================================
Dep. Variable:                  origen   No. Observations:                  406
Model:                         MNLogit   Df Residuals:                      388
Method:                            MLE   Df Model:                           16
Date:               lun., 25 sep. 2023   Pseudo R-squ.:                  0.3514
Time:                          02:25:37   Log-Likelihood:                -242.38
converged:                        False   LL-Null:                       -373.71
Covariance Type:              nonrobust   LLR p-value:                 1.299e-46
==============================================================================
origen=Europa       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Unnamed: 0       -0.0037      0.002     -2.402      0.016      -0.007      -0.001
cv                0.0085      0.019      0.450      0.652      -0.028       0.045
peso             -0.0014      0.002     -0.646      0.518      -0.006       0.003
acel              0.0734      0.095      0.775      0.439      -0.112       0.259
3 cilindros     -21.2687   4.62e+16   -4.6e-16      1.000   -9.06e+16    9.06e+16
4 cilindros       6.7068   1.45e+08   4.62e-08      1.000   -2.85e+08    2.85e+08
5 cilindros      14.4705   1.45e+08   9.95e-08      1.000   -2.85e+08    2.85e+08
6 cilindros       3.9647   1.45e+08   2.73e-08      1.000   -2.85e+08    2.85e+08
8 cilindros     -10.6334   1.46e+08   -7.3e-08      1.000   -2.85e+08    2.85e+08
const            -6.7601   1.45e+08  -4.65e-08      1.000   -2.85e+08    2.85e+08
------------------------------------------------------------------------------
origen=Japón       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
```

```
Unnamed: 0       0.0028      0.002      1.789      0.074      -0.000       0.006
cv               0.0465      0.018      2.525      0.012       0.010       0.083
peso            -0.0116      0.002     -4.760      0.000      -0.016      -0.007
acel             0.1442      0.095      1.512      0.130      -0.043       0.331
3 cilindros     46.9528    1.5e+17   3.12e-16      1.000   -2.95e+17    2.95e+17
4 cilindros     -2.6317   3.73e+07  -7.05e-08      1.000   -7.32e+07    7.32e+07
5 cilindros     -7.4968   3.73e+07  -2.01e-07      1.000   -7.32e+07    7.32e+07
6 cilindros     -3.0375   3.73e+07  -8.14e-08      1.000   -7.32e+07    7.32e+07
8 cilindros    -29.1851   3.73e+07  -7.82e-07      1.000   -7.32e+07    7.32e+07
const            4.6018   3.73e+07   1.23e-07      1.000   -7.32e+07    7.32e+07
==============================================================================
```

En el cuadro anterior podemos observar que el coeficiente de regresión del peso en los automóviles es significativo y negativo, lo que significa que un menor peso en un vehículo incrementa la probabilidad de que el automóvil sea de origen Europeo o Japonés con Respecto a ser de origen Americano, de esta manera se pueden interpretar tambien los demas coeficientes.

1.2.4. Predicciones

Debido a que la clase de referencia es Estados Unidos, las probabilidades de pertenecer al origen de Japón o Europa son con respecto a Estados Unidos, ya que en la sección anterior se obtuvieron los coeficientes de regresión, se muestran a continuación como se calcula la probabilidad de pertenecer a cada origen:

$$\text{pr(Europa)}= \frac{e^{0.0085\times cv-0.0014\times peso+0.0734\times acel-21.27\times 3cilindros+6.7\times 4cilindros+14.47\times 5cilindros+3.96\times 6cilindros-10.63\times 8cilindros-6.76}}{1+e^{0.0085\times cv-0.0014\times peso+0.0734\times acel-21.27\times 3cilindros+6.7\times 4cilindros+14.47\times 5cilindros+3.96\times 6cilindros-10.63\times 8cilindros-6.76}}$$

$$\text{pr(Japón)}= \frac{e^{0.047\times cv-0.001\times peso+0.14\times acel+46.95\times 3cilindros-2.63\times 4cilindros-7.5\times 5cilindros-3\times 6cilindros-29.19\times 8cilindros+4.6}}{1+e^{0.047\times cv-0.001\times peso+0.14\times acel+46.95\times 3cilindros-2.63\times 4cilindros-7.5\times 5cilindros-3\times 6cilindros-29.19\times 8cilindros+4.6}}$$

$$\text{pr(EE.UU.)}= 1-pr(Europa)-pr(Japon)$$

Para poder generar predicciones, se creará una función en Python que nos de una predicción sobre cual es el origen geografico mas posible de un nuevo automovil, la función se muestra a continuación.

```python
from math import e

def prediccion(cv,peso,acel,cilindro):
  if cilindro == '8 cilindros':
    expo= e**(cv*0.0085-0.0014*peso+0.0734*acel-10.6334-6.7601)
    europa=expo/(1+expo)
  elif cilindro == '6 cilindros':
    expo= e**(cv*0.0085-0.0014*peso+0.0734*acel+3.9647-6.7601)
```

```
    europa=expo/(1+expo)
  elif cilindro == '5 cilindros':
    expo= e**(cv*0.0085-0.0014*peso+0.0734*acel+14.4705-6.7601)
    europa=expo/(1+expo)
  elif cilindro == '4 cilindros':
    expo= e**(cv*0.0085-0.0014*peso+0.0734*acel+6.7086-6.7601)
    europa=expo/(1+expo)
  else:
    expo= e**(cv*0.0085-0.0014*peso+0.0734*acel-21.2687-6.7601)
    europa=expo/(1+expo)

#japón
  if cilindro=='8 cilindros':
    expo2=e**(cv*0.0465-0.0116*peso+0.1442*acel-29.1851+4.6018)
    expo2=expo2/(1+expo2)
  elif cilindro=='6 cilindros':
    expo2=e**(cv*0.0465-0.0116*peso+0.1442*acel-3.0375+4.6018)
    expo2=expo2/(1+expo2)
  elif cilindro=='5 cilindros':
    expo2=e**(cv*0.0465-0.0116*peso+0.1442*acel-7.4968+4.6018)
    expo2=expo2/(1+expo2)
  elif cilindro=='4 cilindros':
    expo2=e**(cv*0.0465-0.0116*peso+0.1442*acel-2.6317+4.6018)
    expo2=expo2/(1+expo2)
  else:
    expo2=e**(cv*0.0465-0.0116*peso+0.1442*acel+46.9528+4.6018)
    expo2=expo2/(1+expo2)

  eeuu=1-(expo+expo2)

  if eeuu>expo and eeuu>expo2:
    return 'EE.UU.'
  elif expo2>expo:
    return 'Japón'
  else:
```

```
    return 'Europa'

print(prediccion(130,1168,12,'8 cilindros'))
```

```
EE.UU.
```

Ahora se realizarán predicciones para los siguientes 5 nuevos registros de automóviles:

```
autos= ['Automovil 1','Automovil 2','Automovil 3','Automovil 4','Automovil 5']
pot= [220,71,59,61,108]
p= [1225,856,605,498,905]
ac= [11,14,19,17,16]
cil= ['8 cilindros','4 cilindros','4 cilindros','5 cilindros','6 cilindros']

print(pd.DataFrame({'Auto':autos,'cv':pot,'peso':p,'aceleracion':ac,'cilindro':cil}))
```

```
         Auto   cv  peso  aceleracion     cilindro
0  Automovil 1  220  1225           11  8 cilindros
1  Automovil 2   71   856           14  4 cilindros
2  Automovil 3   59   605           19  4 cilindros
3  Automovil 4   61   498           17  5 cilindros
4  Automovil 5  108   905           16  6 cilindros
```

A continuación, se muestra la clase predicha para cada uno de estos nuevos 5 automóviles, la cual se ha predicho mediante el modelo de regresión logística multinomial estimado.

```
pred=[]

for i in range(len(autos)):
  pred.append(prediccion(pot[i],p[i],ac[i],cil[i]))

print(pd.DataFrame({'Automóvil':autos,'Origen Predicho':pred}))
```

```
     Automóvil Origen Predicho
0  Automovil 1          EE.UU.
1  Automovil 2          Europa
2  Automovil 3          Europa
3  Automovil 4          Europa
```

```
4  Automovil 5          EE.UU.
```

2. Regresión Logística Multinomial en R

Se tiene interés en predecir el origen geográfico de automóviles tomando en cuenta su potencia, peso, aceleración, número de cilindros y si el automóvil fue seleccionado o no al pasar por un filtro, pero previo a realizar una estimación de un modelo de regresión logística multinomial con el origen geográfico como variable dependiente, se analizará que variables son las que en verdad tienen sentido que se toman en cuenta para la estimación del modelo.

2.1. Análisis exploratorio

Se importará la base de datos que contiene las variables necesarias para realizar el ajuste del modelo de elección múltiple, también se observará cuáles son las variables que contiene la base de datos y cuantos registros tiene.

```
library(readxl)


autos <- read_excel('automoviles.xlsx')
str(autos)
```

```
tibble [406 x 6] (S3: tbl_df/tbl/data.frame)
 $ Potencia   : num [1:406] 130 165 150 150 140 198 220 215 225 190 ...
 $ peso       : num [1:406] 1168 1231 1145 1144 1149 ...
 $ aceleracion: num [1:406] 12 12 11 12 11 10 9 9 10 9 ...
 $ origen     : chr [1:406] "EE.UU." "EE.UU." "EE.UU." "EE.UU." ...
 $ cilindro   : chr [1:406] "8 cilindros" "8 cilindros" "8 cilindros" "8 cilindros" ...
 $ filtro     : chr [1:406] "No Seleccionado" "No Seleccionado" "No Seleccionado" "No Seleccionado" ...
```

2.1.1. Gestión de valores nulos

A continuación, se mostrará si las variables que se utilizarán para la estimación del modelo logístico multinomial contienen valores nulos.

```
sapply(autos, function(x) sum(is.na(x)))
```

```
   Potencia        peso aceleracion      origen    cilindro      filtro 
          6           0           0           1           1           2 
```

Como podemos observar en el resultado anterior, las variables potencia, origen, cilindro y filtro contienen valores nulos.

A continuación, se muestra como imputamos los valores perdidos de la variable potencia, mediante el método predictive mean machine, el cual es de los más polivalentes debido a que rellena los datos faltantes con valores presentes en el conjunto de datos y eso reduce el sesgo y la probabilidad de tener valores anómalos.

```
library(mice)
```

```
Warning: package 'mice' was built under R version 4.3.1
```

```
Attaching package: 'mice'
```

```
The following object is masked from 'package:stats':

    filter
```

```
The following objects are masked from 'package:base':

    cbind, rbind
```

```
imputacion <- mice(autos, m=5, meth = 'pmm', seed = 13)
```

```
 iter imp variable
  1   1  Potencia
  1   2  Potencia
  1   3  Potencia
  1   4  Potencia
  1   5  Potencia
  2   1  Potencia
  2   2  Potencia
  2   3  Potencia
  2   4  Potencia
  2   5  Potencia
  3   1  Potencia
  3   2  Potencia
  3   3  Potencia
  3   4  Potencia
  3   5  Potencia
  4   1  Potencia
```

```
4   2  Potencia
4   3  Potencia
4   4  Potencia
4   5  Potencia
5   1  Potencia
5   2  Potencia
5   3  Potencia
5   4  Potencia
5   5  Potencia
```

```
Warning: Number of logged events: 3
```

```
autos<- complete(imputacion, 1)
```

Ahora se muestra como se imputan por la moda los valores faltantes en las variables cualitativas que poseen ese tipo de valores.

```
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

autos$filtro<-ifelse(is.na(autos$filtro),getmode(autos$filtro),autos$filtro)
autos$origen<-ifelse(is.na(autos$origen),getmode(autos$origen),autos$origen)
autos$cilindro<-ifelse(is.na(autos$cilindro),getmode(autos$cilindro),
                       autos$cilindro)
```

2.1.2. Análisis descriptivo

Previo a la estimación del modelo de elección múltiple, es muy importante realizar un análisis descriptivo bivariado que nos permita tener una intuición de la relación que puede tener la variable origen con las variables independientes.

A continuación empezaremos analizando la relación entre la variable origen y la variable número de cilindros.

```
library('gmodels')
attach(autos)
CrossTable(origen,cilindro,prop.chisq=FALSE,prop.c=FALSE,prop.r=FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  406
```

	cilindro					
origen	3 cilindros	4 cilindros	5 cilindros	6 cilindros	8 cilindros	Row Total
EE.UU.	0	73	0	74	107	254
	0.000	0.180	0.000	0.182	0.264	
Europa	0	66	3	4	0	73
	0.000	0.163	0.007	0.010	0.000	
Japón	4	69	0	6	0	79
	0.010	0.170	0.000	0.015	0.000	
Column Total	4	208	3	84	107	406

Como podemos observar en la tabla anterior, parece ser que los automóviles de 4 cilindros son de los que se tienen cantidades más o menos similares en cada una de las 3 posibles regiones de origen (Estados Unidos, Europa o Japón), y que sí, un automóvil es de 8 cilindros prácticamente es seguro que su origen sea Estados unidos; sin embargo, los automóviles de Estados Unidos parece ser que se distribuyen de manera uniforme en cuanto a los que son de 4 y 6 cilindros. En conclusión, se nota que esta variable puede aportarnos mucha información para el modelo logístico que se desea estimar, por lo que esta variable será incluida como variable explicativa. Ahora se analizará la relación entre la variable origen y la variable derivada (Filtro).

```
CrossTable(origen,filtro,prop.chisq=FALSE,prop.c=FALSE,prop.r=FALSE)
```

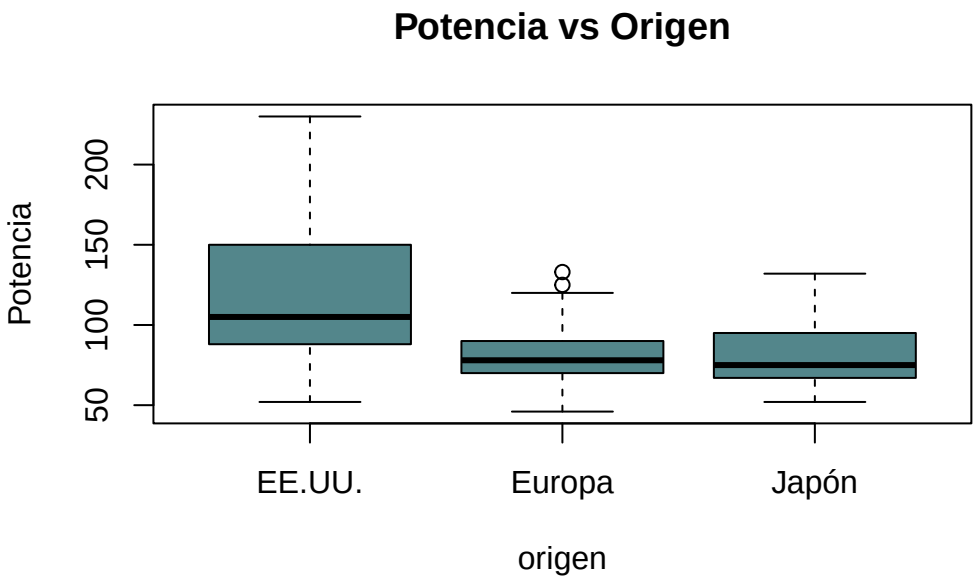
```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  406


             | filtro
      origen | No Seleccionado |    Seleccionado |       Row Total |
-------------|-----------------|-----------------|-----------------|
      EE.UU. |             107 |             147 |             254 |
             |           0.264 |           0.362 |                 |
-------------|-----------------|-----------------|-----------------|
      Europa |               0 |              73 |              73 |
             |           0.000 |           0.180 |                 |
-------------|-----------------|-----------------|-----------------|
       Japón |               0 |              79 |              79 |
             |           0.000 |           0.195 |                 |
-------------|-----------------|-----------------|-----------------|
Column Total |             107 |             299 |             406 |
-------------|-----------------|-----------------|-----------------|
```

Como podemos observar en el resultado anterior, parece ser que los automóviles de origen Europa y Japón son seleccionados prácticamente el 100% mientras que los de origen en Estados unidos son seleccionados aproximadamente el 50% (esto puede restar potencialidad a la capacidad predictiva del modelo), es por esta razón que la variable filtro (derivada) se considera que no vale la pena incluirla en nuestro modelo.

A continuación seguiremos con el análisis de la variable potencia, ya que es una variable métrica, en la cual se analizará que tanto varían las potencias de los automóviles con respecto a los orígenes de los mismos, estas

diferencias se muestran en la siguiente figura.

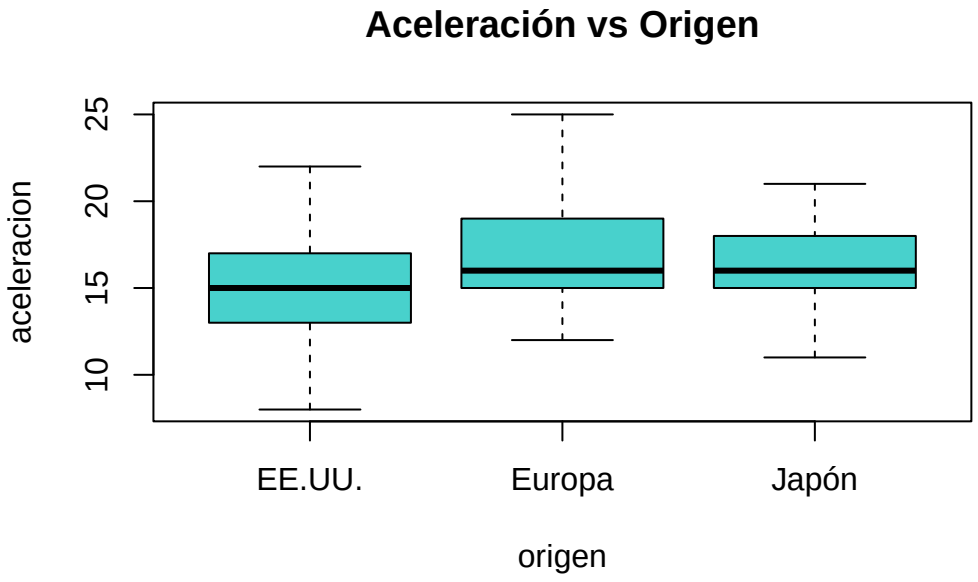
```
boxplot(Potencia~origen, main='Potencia vs Origen', col='cadetblue4')
```



Como podemos observar en la figura anterior, parece ser que las potencias medias de los vehículos con respecto a su origen varían, aunque la mayor diferencia se observa con los automóviles de Estados Unidos, los cuales difieren y varían más con respecto a los de Europa y Japón, que es poco lo que difieren y varían entre ellos, por esta razón se ha decidido tomar en cuenta esta variable como variable explicativa.

Ahora se analizará si vale la pena incluir la variable aceleración como variable explicativa en nuestro modelo de respuesta cualitativa, esto se analizará de la misma manera que con la variable potencia. A continuación se muestra el respectivo gráfico de la aceleración con respecto a su origen.

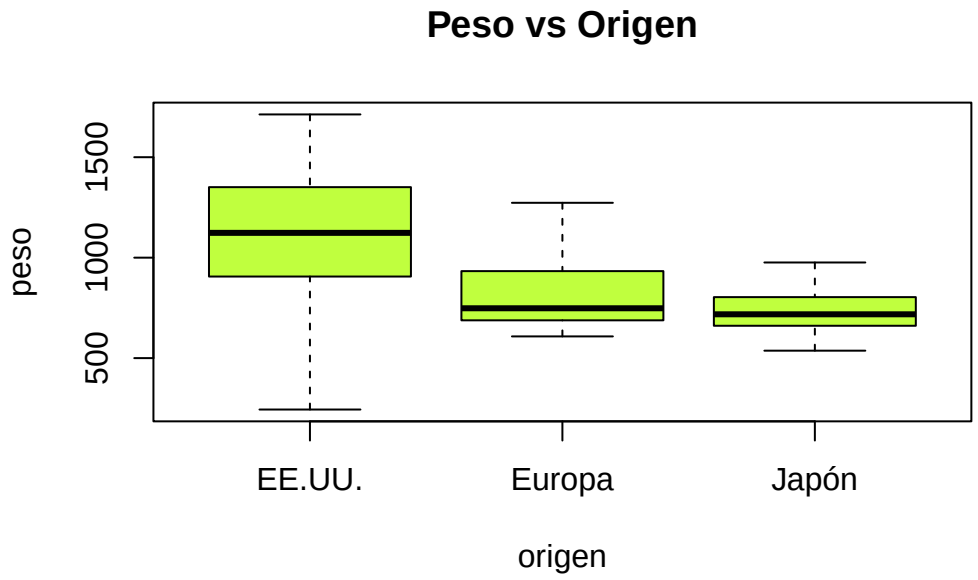
```
boxplot(aceleracion~origen, main='Aceleración vs Origen', col='mediumturquoise')
```



Como podemos observar en la figura anterior, se observa que la diferencia en la aceleración medio de los automóviles con origen en Europa difiere aproximadamente igual con los de origen en Japón y Estados Unidos, y la diferencia entre la aceleración media es más grande entre Japón y Estados unidos que las anteriormente mencionadas, también cabe mencionar que en Europa es donde se da la aceleración promedio más alta en los automóviles, debido a estas diferencias descritas se decide por tomar en cuenta la variable aceleración para nuestro modelo de respuesta múltiple.

Ahora analizaremos la discriminación del origen con respecto al peso de los vehículos, para realizar este análisis se muestra a continuación, un boxplot del peso de los automóviles con respecto a su origen.

```
boxplot(peso~origen, main='Peso vs Origen', col='olivedrab1')
```



Como podemos observar en la figura anterior, los pesos medios difieren con respecto a su origen, es decir, el origen de los automóviles tiene influencia en su peso, si miramos los gráficos de Estados Unidos y Japón ni siquiera se traslapan, por ello se decide que si vale la pena agregar la variable peso a nuestro modelo de respuesta múltiple.

2.2. Estimación del modelo de elección múltiple

A continuación se cargan todos los paquetes necesarios para la estimación del modelo de elección múltiple.

```
library(foreign)
library(nnet)
```

```
Warning: package 'nnet' was built under R version 4.3.1
```

```
library(stargazer)
```

```
Please cite as:

 Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

 R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.3.1
```

```
library(gmodels)
library(stargazer)
```

Para estimar el modelo se muestra a continuación un vistazo a los primeros 6 registros de la base de datos que contiene las variables.

```
head(autos)
```

```
  Potencia peso aceleracion origen    cilindro          filtro
1      130 1168          12 EE.UU. 8 cilindros No Seleccionado
2      165 1231          12 EE.UU. 8 cilindros No Seleccionado
3      150 1145          11 EE.UU. 8 cilindros No Seleccionado
4      150 1144          12 EE.UU. 8 cilindros No Seleccionado
5      140 1149          11 EE.UU. 8 cilindros No Seleccionado
6      198 1447          10 EE.UU. 8 cilindros No Seleccionado
```

Ahora se mostrarán los niveles de o categorías de las variables no métricas que se utilizarán para realizar la estimación del modelo de elección discreta.

```
#Origenes posible
unique(autos$origen)
```

```
[1] "EE.UU." "Europa" "Japón"
```

```
#posibles números de cilindro
unique(autos$cilindro)
```

```
[1] "8 cilindros" "4 cilindros" "6 cilindros" "3 cilindros" "5 cilindros"
```

Como podemos observar, hay 3 orígenes posible y 5 posibles números de cilindros.

Después de haber intuido la relación entre la variable dependiente y las independientes, nos encontramos en condición de estimar el modelo logístico multinomial, en el cual se tomará a Japón como el origen geográfico de referencia respecto al cual se va a interpretar los coeficientes de los otros dos factores.

```
library(stats)
autos$origen <- as.factor(autos$origen)
autos$origen2<- relevel(autos$origen, ref = "EE.UU.")
```

A continuación se muestra la estimación del modelo logístico multinomial en R.

```
multinomial <- multinom(origen2 ~ Potencia + peso + aceleracion + cilindro,
data=autos)
```

```
# weights:  27 (16 variable)
initial  value 446.036589
iter  10 value 268.516355
iter  20 value 248.122370
iter  30 value 248.038148
final  value 248.037711
converged
```

```
summary(multinomial)
```

```
Call:
multinom(formula = origen2 ~ Potencia + peso + aceleracion +
    cilindro, data = autos)

Coefficients:
       (Intercept)   Potencia         peso aceleracion cilindro4 cilindros
Europa   -8.403523 0.03437585 -0.004260876   0.1863658            5.807745
Japón    21.248939 0.05171190 -0.011606279   0.1592519          -19.188606
       cilindro5 cilindros cilindro6 cilindros cilindro8 cilindros
Europa           27.211847            3.441231           -13.37253
Japón            -6.376202          -19.783858           -39.67347

Std. Errors:
       (Intercept)  Potencia        peso aceleracion cilindro4 cilindros
Europa  0.03749867 0.0141322 0.002054483  0.06081274           0.3213209
Japón   0.03481796 0.0156762 0.002383069  0.06439436           0.2988285
       cilindro5 cilindros cilindro6 cilindros cilindro8 cilindros
Europa        3.973585e-06           0.2844551        9.074763e-09
Japón         3.973530e-06           0.2646285        1.078183e-09

Residual Deviance: 496.0754
AIC: 528.0754
```

Como podemos observar en el resultado anterior, ya se cuenta con la estimación del modelo logístico multinomial, en el que nos mostró cada estimación de los parámetros con su respectivo error estándar, también se cuenta con el valor de la función de máxima verosimilitud de los residuales multiplicada por -2, ahora mediante esta información que nos brinda la estimación del modelo se calculará la significatividad global del modelo y la significatividad individual para cada coeficiente, ya que estos cálculos son necesarios para la interpretación del modelo.

2.2.1. Contraste de hipótesis para el modelo estimado

Para poder calcular el ratio de verosimilitud se estimará el mismo modelo que se calculó anteriormente, pero solo con la constante, y se guardará la deviance del modelo estimado ($-2LL$) y se calculará la diferencia entre ambas, ya que está diferencia sigue una distribución X^2. A continuación se muestra la realización de estos cálculos.

```
# Estimamos el modelo solo con la constante
multi0 <- multinom(origen2 ~ 1, data=autos)
```

```
# weights:  6 (2 variable)
initial  value 446.036589
iter  10 value 373.706559
iter  10 value 373.706559
iter  10 value 373.706559
final  value 373.706559
converged
```

```
# Calculamos el estadístico chi cuadrado como
# diferencia de sus -2LL (deviance)

chi2<-multi0$deviance-multinomial$deviance
df.chi2<-multinomial$edf-multi0$edf
Sig.chi2<-1-pchisq(chi2,df.chi2)
print(cbind(chi2,df.chi2,Sig.chi2))
```

```
         chi2 df.chi2 Sig.chi2
[1,] 251.3377      14        0
```

Como podemos observar en el resultado anterior, el valor del estadístico de significatividad global es prácticamente cero, lo que significa que al menos un coeficiente está teniendo una influencia significativa para ayudar a predecir el origen geográfico de los automóviles, es por ello que se considera que las variables que se

han determinado como independientes después del análisis descriptivo de los datos tienen peso al momento de tener influencia en el origen de los automóviles.

Para calcular la significatividad individual de los parámetros, es suficiente con calcular la división entre los parámetros con su respectivo error estándar y se compara este resultado con el valor crítico de la distribución normal. Esto se realizará a continuación, mediante la función stargazer.

```
stargazer(multinomial, type = "text")
```

```
================================================
                        Dependent variable:
                    ----------------------------
                        Europa          Japón
                          (1)             (2)
------------------------------------------------
Potencia                0.034**        0.052***
                        (0.014)         (0.016)

peso                    -0.004**       -0.012***
                        (0.002)         (0.002)

aceleracion             0.186***        0.159**
                        (0.061)         (0.064)

cilindro4 cilindros     5.808***      -19.189***
                        (0.321)         (0.299)

cilindro5 cilindros    27.212***       -6.376***
                       (0.00000)       (0.00000)

cilindro6 cilindros     3.441***      -19.784***
                        (0.284)         (0.265)

cilindro8 cilindros    -13.373***     -39.673***
                        (0.000)         (0.000)

Constant               -8.404***       21.249***
```

```
                        (0.037)        (0.035)

-----------------------------------------------
Akaike Inf. Crit.       528.075        528.075
===============================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Como podemos observar en el resultado anterior, prácticamente todas las variables aportan información para poder calcular la probabilidad de que un automóvil pertenezca a un determinado origen, y mediante esa probabilidad se realiza la predicción sobre a cuál de los 3 posibles orígenes es más plausible que pertenezca un determinado automóvil, es por esta razón que el modelo seguirá de la misma manera que sé ha estimado.

2.2.2. Interpretación de los coeficientes de regresión

En el cuadro anterior podemos observar que el coeficiente de regresión del peso en los automóviles es significativo y positivo, lo que significa que un mayor peso en un vehículo incrementa la probabilidad de que el automóvil sea de origen Europeo o Japonés con Respecto a ser de origen Americano.

A continuación se calcularán los risk ratios, los cuales son un equivalente a los odd ratios que se estudian en la regresión logística binaria, esto con el fin de analizar e interpretar la contribución relativa de cada variable independiente.

```
multi1.rrr = exp(coef(multinomial))

stargazer(multinomial, type="text", coef=list(multi1.rrr), p.auto=FALSE)
```

```
============================================================
                            Dependent variable:
                    ---------------------------------------
                           Europa              Japón
                             (1)                (2)
------------------------------------------------------------
Potencia                   1.035**            1.053***
                           (0.014)            (0.016)

peso                       0.996**            0.988***
                           (0.002)            (0.002)
```

```
aceleracion                1.205***              1.173**
                           (0.061)               (0.064)

cilindro4 cilindros       332.868***             0.000***
                           (0.321)               (0.299)

cilindro5 cilindros 657,589,555,168.000***       0.002***
                          (0.00000)             (0.00000)

cilindro6 cilindros       31.225***              0.000***
                           (0.284)               (0.265)

cilindro8 cilindros       0.00000***             0.000***
                           (0.000)               (0.000)

Constant                  0.0002***        1,691,596,995.000***
                           (0.037)               (0.035)

-----------------------------------------------------------------
Akaike Inf. Crit.          528.075               528.075
=================================================================
Note:                                *p<0.1; **p<0.05; ***p<0.01
```

Podemos observar que en el cuadro anterior, el mayor impacto sobre el origen geográfico de un automóvil, lo tiene el nivel de la variable cilindro, que corresponde a automóviles con 4 cilindros, esto significa que es 0.62 veces más probables que un país pertenezca a Europa o Japón con respecto a Estados Unidos (categoría de referencia).

2.2.3. Evaluación del ajuste del modelo

Para evaluar el ajuste del modelo, se calculará el R^2 de Mc Fadden el cual es intento de equivalencia con respecto al R^2 de la regresión lineal, este estadístico se construye como la diferencia entre la deviance del modelo constante y el modelo estimado entre la deviance del modelo constante, si el modelo estimado minimiza esa función de máxima verosimilitud es porque los coeficientes están teniendo una significativa influencia para predecir el origen geográfico de los automóviles, y la diferencia en el numerador será mínima si esto sucede, y entonces el valor del R^2 de Mc Fadden será mayor que cero, a continuación, para nuestro modelo de elección múltiple se muestra este cálculo descrito.

```
R2MF<-(multi0$deviance-multinomial$deviance)/multi0$deviance;R2MF
```

```
[1] 0.3362768
```

Como podemos observar, el valor del R^2 de Mc Fadden es de 0.34 aproximadamente, lo que significa que el modelo estimado ayuda a predecir con menos errores los orígenes geográficos de los automóviles.

2.2.4. Capacidad discriminante del modelo

Para evaluar la capacidad discriminante del modelo se construirá una matriz de confusión, la cual es una tabla cruzada entre los orígenes reales de los automóviles y los orígenes predichos por el modelo de regresión logística multinomial, a continuación se muestra la construcción de esta matriz.

```
library(nnet)
predicciones <- predict(multinomial, autos, type="class")
reales <- autos$origen
CrossTable(predicciones,reales,prop.chisq=FALSE,prop.c=FALSE,prop.r=FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  406


             | reales
predicciones |    EE.UU. |    Europa |     Japón | Row Total |
-------------|-----------|-----------|-----------|-----------|
      EE.UU. |       223 |        27 |        22 |       272 |
             |     0.549 |     0.067 |     0.054 |           |
-------------|-----------|-----------|-----------|-----------|
      Europa |        11 |        13 |         5 |        29 |
             |     0.027 |     0.032 |     0.012 |           |
-------------|-----------|-----------|-----------|-----------|
```

```
       Japón |        20 |        33 |        52 |       105 |
             |     0.049 |     0.081 |     0.128 |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |       254 |        73 |        79 |       406 |
-------------|-----------|-----------|-----------|-----------|
```

Como podemos observar en la matriz de confusión, es una cantidad muy elevada de predicciones acertadas con respecto al origen de Estados Unidos es decir que los automóviles que tengan el perfil de Estados Unidos tienen una probabilidad del 82% de clasificarse correctamente, después siguen los orígenes de Europa y Japón los cuales también tienen un buen acierto, pero con menor efectividad que en Estados Unidos, pero en general se tiene una capacidad predictiva en el modelo de elección múltiple de aproximadamente el 71% de efectividad, lo cual significa que se tiene un modelo bien ajustado y se espera que esa capacidad predictiva se aproveche en los casos donde se requiera, por ejemplo alguien que se dedica a la venta de Vehículos pudiera estar muy interesado es predecir el origen, con el fin de cotizar el precio de los vehículos.

2.2.5. Predicciones

Uno de los objetivos de los modelos estadísticos es realizar predicciones en virtud del perfil de los objetos que se van a clasificar en un determinado grupo, para este problema, los objetos son los automóviles, y mediante el modelo de elección multiple que se estimó anteriormente se realizarán predicciones para 5 nuevos registros de automóviles de los cuales se quiere predecir el origen geográfico y cuyos datos de cada uno de estos nuevos automóviles muestran a continuación en el siguiente data frame.

```r
autos<-paste('Auto nuevo',1:5)
pot<- c(220,71,59,61,108)
p <- c(1225,856,605,698,905)
ac <- c(11,14,19,17,16)
cil <- c('8 cilindros','4 cilindros','4 cilindros','5 cilindros','6 cilindros')

data.frame(autos,Potencia=pot,peso=p,aceleracion=ac,cilindro=cil)
```

```
        autos Potencia peso aceleracion    cilindro
1 Auto nuevo 1      220 1225          11 8 cilindros
2 Auto nuevo 2       71  856          14 4 cilindros
3 Auto nuevo 3       59  605          19 4 cilindros
4 Auto nuevo 4       61  698          17 5 cilindros
```

```
5 Auto nuevo 5      108  905          16 6 cilindros
```

A continuación se muestra el origen predicho para cada uno de los 5 nuevos automóviles más resientes de los cuales se tiene registro, y como podemos observar, los automóviles de origen Estadounidense parecen ser los más frecuentes.

```
AutosNuevos <- data.frame(Potencia=pot,peso=p,aceleracion=ac,cilindro=cil)

prediccion <- predict(multinomial,newdata =AutosNuevos)

OrigenesPredichos<-data.frame(Autos=autos, 'Origen Predicho'=prediccion)

OrigenesPredichos
```

```
         Autos Origen.Predicho
1 Auto nuevo 1          EE.UU.
2 Auto nuevo 2          EE.UU.
3 Auto nuevo 3           Japón
4 Auto nuevo 4          Europa
5 Auto nuevo 5          EE.UU.
```

3. Regresión Logística Multinomial en IBM SPSS

3.1. Análisis exploratorio

En spss no se sustituirán los valores nulos, debido a que spss detecta cuales son los valores nulos que contiene nuestra base de datos y realiza los analisis sin tomarlos en cuenta.

3.1.1 Análisis descriptivo

Para poder determinar cuáles variables independientes valen la pena agregar a nuestro modelo de elección múltiple se analizará de manera bi variada la relación entre la variable dependiente (origen) con las variables independientes, mediante tablas de contingencia se analizará la relación entre la variable dependiente (origen) y las variables no métricas (número de cilindros y derivada) y mediante gráficos de cajas se analizará la relación entre las variables independientes métricas (peso, aceleración y potencia) y la variable dependiente origen, y estos gráficos se harán por categorías de la variable origen. A continuación empezaremos analizando la relación entre la variable origen y la variable número de cilindros.

Tabla cruzada (País de origen y Número de cilindros)

Recuento

		Número de cilindros					
		3 cilindros	4 cilindros	5 cilindros	6 cilindros	8 cilindros	Total
País de origen	EE.UU.	0	72	0	74	107	253
	Europa	0	66	3	4	0	73
	Japón	4	69	0	6	0	79
Total		4	207	3	84	107	405

Como podemos observar en la tabla anterior, parece ser que los automóviles de 4 cilindros son de los que se tienen cantidades más o menos similares en cada una de las 3 posibles regiones de origen (Estados Unidos, Europa o Japón), y que sí, un automóvil es de 8 cilindros prácticamente es seguro que su origen sea Estados unidos; sin embargo, los automóviles de Estados Unidos parece ser que se distribuyen de manera uniforme en cuanto a los que son de 4 y 6 cilindros. En conclusión, se nota que esta variable puede aportarnos mucha información para el modelo logístico que se desea estimar, por lo que esta variable será incluida como variable explicativa. Ahora se analizará la relación entre la variable origen y la variable derivada (Filtro).

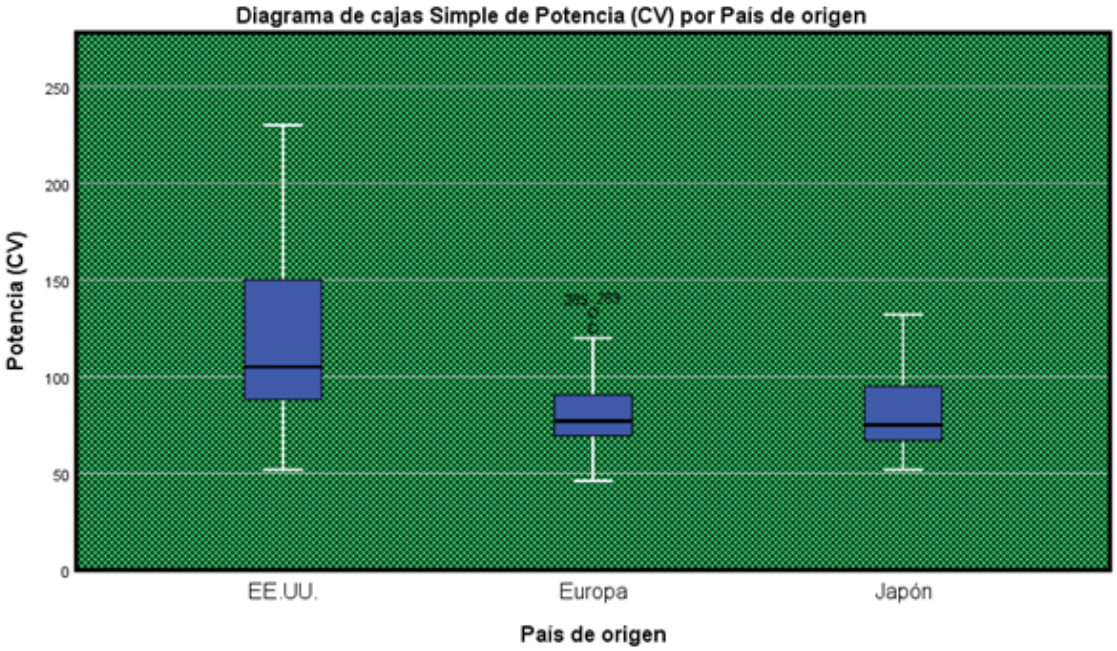
Tabla cruzada (País de origen y FILTRO)

Recuento

		FILTRO		
		No Seleccionado	Seleccionado	Total
País de origen	EE.UU.	107	146	253
	Europa	0	73	73
	Japón	0	78	78
Total		107	297	404

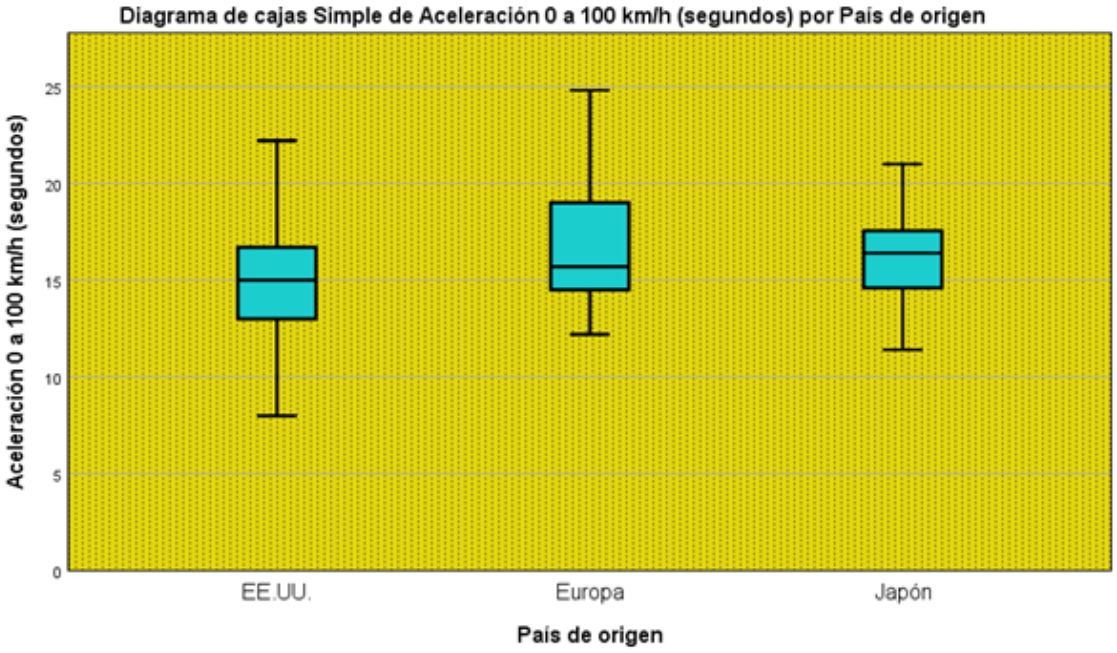
Como podemos observar en el resultado anterior, parece ser que los automóviles de origen Europa y Japón son seleccionados prácticamente el 100% mientras que los de origen en Estados unidos son seleccionados aproximadamente el 50% (esto puede restar potencialidad a la capacidad predictiva del modelo), es por esta razón que la variable filtro (derivada) se considera que no vale la pena incluirla en nuestro modelo.

A continuación seguiremos con el análisis de la variable potencia, ya que es una variable métrica, en la cual se analizará que tanto varían las potencias de los automóviles con respecto a los orígenes de los mismos, estas diferencias se muestran en la siguiente figura.



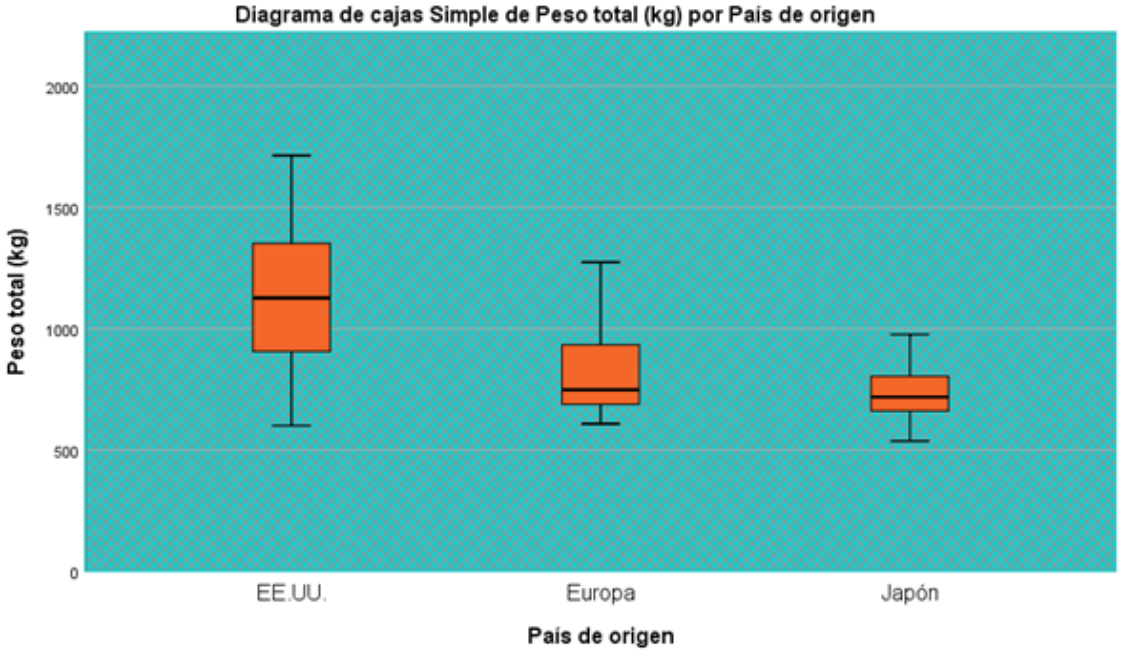
Como podemos observar en la figura anterior, parece ser que las potencias medias de los vehículos con respecto a su origen varían, aunque la mayor diferencia se observa con los automóviles de Estados Unidos, los cuales difieren y varían más con respecto a los de Europa y Japón, que es poco lo que difieren y varían entre ellos, por esta razón se ha decidido tomar en cuenta esta variable como variable explicativa.

Ahora se analizará si vale la pena incluir la variable aceleración como variable explicativa en nuestro modelo de respuesta cualitativa, esto se analizará de la misma manera que con la variable potencia. A continuación se muestra el respectivo gráfico de la aceleración con respecto a su origen.



Como podemos observar en la figura anterior, se observa que la diferencia en la aceleración medio de los automóviles con origen en Europa difiere aproximadamente igual con los de origen en Japón y Estados Unidos, y la diferencia entre la aceleración media es más grande entre Japón y Estados unidos que las anteriormente mencionadas, también cabe mencionar que en Europa es donde se da la aceleración promedio más alta en los automóviles, debido a estas diferencias descritas se decide por tomar en cuenta la variable aceleración para nuestro modelo de respuesta múltiple.

Ahora analizaremos la discriminación del origen con respecto al peso de los vehículos, para realizar este análisis se muestra a continuación, un boxplot del peso de los automóviles con respecto a su origen.



Como podemos observar en la figura anterior, los pesos medios difieren con respecto a su origen, es decir, el origen de los automóviles tiene influencia en su peso, si miramos los gráficos de Estados Unidos y Japón ni siquiera se traslapan, por ello se decide que si vale la pena agregar la variable peso a nuestro modelo de respuesta múltiple.

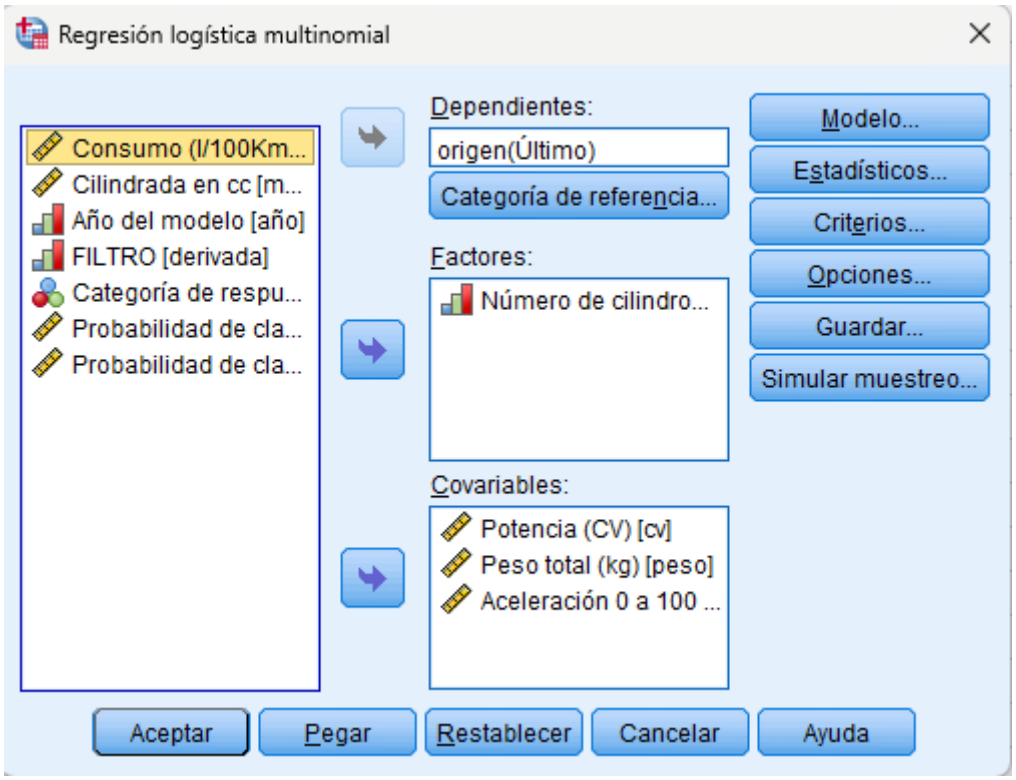
3.2. Estimación del modelo de elección múltiple

A continuación, se muestra en la siguiente tabla, el resumen del procesamiento de casos para la estimación del modelo de regresión logística multinomial, en el cual se muestra que hay 7 valores perdidos, y SPSS identifica cuales son los valores perdidos y no los toma en cuenta para el análisis, es por eso que acá se decidió no gestionar los valores nulos, y mostraros los casos que hay para cada categoría de las variables no métricas que participarán en el modelo.

Resumen de procesamiento de casos

		N	Porcentaje marginal
País de origen	EE.UU.	249	62.4%
	Europa	71	17.8%
	Japón	79	19.8%
Número de cilindros	3 cilindros	4	1.0%
	4 cilindros	202	50.6%
	5 cilindros	3	0.8%
	6 cilindros	83	20.8%
	8 cilindros	107	26.8%
Válidos		399	100.0%
Perdidos		7	
Total		406	
Subpoblación		397[a]	

A continuación, se muestra en el siguiente cuadro el ajuste del modelo de regresión logística multinomial en IBM SPSS, en el cual especificamos que el origen geográfico será nuestra variable dependiente y que las variables potencia, aceleración, peso y número de cilindros (factor) serán las variables independientes o explicativas de nuestro modelo de regresión logística multinomial.



3.2.1 Contraste de hipótesis para el modelo estimado

A continuación, en el siguiente cuadro se muestra la prueba de razón de máxima verosimilitud para el modelo estimado.

Modelo	Criterios de ajuste de modelo: Logaritmo de la verosimilitud -2	Pruebas de la razón de verosimilitud: Chi-cuadrado	gl	Sig.
Sólo intersección	735.826			
Final	472.104	263.723	14	.000

Como podemos observar en el cuadro anterior, el resultado del valor p (sig) es prácticamente cero lo que significa que los coeficientes si están ejerciendo influencia en nuestro modelo para poder predecir el origen geográfico de los automóviles, el valor de la -2 veces la función de máxima verosimilitud con los coeficientes es mucho menor que la que se realizo solo con el termino constante, es decir, hay una drástica diferencia, eso significa que la función de máxima verosimilitud se minimiza cuando se tienen en el modelo las variables explicativas que se han considerado que nos ayudarán para predecir con un menor grado de error el origen geográfico de un determinado automóvil.

A continuación, en el siguiente cuadro, se muestra el contraste de significatividad individual para determinar el peso de cada variable (coeficiente) para determinar un origen predicho a un determinado automóvil.

Pruebas de la razón de verosimilitud

Efecto	Criterios de ajuste de modelo: Logaritmo de la verosimilitud -2 de modelo reducido	Pruebas de la razón de verosimilitud: Chi-cuadrado	gl	Sig.
Intersección	472.104[a]	.000	0	.
Potencia (CV)	486.224	14.120	2	.001
Peso total (kg)	513.231	41.127	2	.000
Aceleración 0 a 100 km/h (segundos)	476.408	4.304	2	.116
Número de cilindros	521.631	49.527	8	.000

Para la realización de esta prueba se calcula el valor de la deviance ($-2LL$) con y sin el coeficiente, para determinar si tiene influencia significativa en el modelo de elección discreta mediante la diferencia entre

las dos estimaciones, y como se puede observar en el cuadro anterior, parece ser que todas las variables explicativas tienen influencia a la hora de calcular la probabilidad de que un automóvil pertenezca a un destino geográfico, excepto la variable aceleración, la cual parece ser que no tiene mucha influencia.

3.2.2. Interpretación de los coeficientes de regresión

A continuación, mediante spss se muestra la estimación de los coeficientes de regresión.

Estimaciones de parámetro

País de origen[a]		B	Desv. Error	Wald	gl	Sig.	Exp(B)	95% de intervalo de confianza para Exp(B)	
								Límite inferior	Límite superior
EE.UU.	Intersección	15.533	571.755	.001	1	.978			
	Potencia (CV)	-.074	.020	13.294	1	.000	.929	.893	.966
	Peso total (kg)	.015	.003	31.042	1	.000	1.015	1.010	1.021
	Aceleración 0 a 100 km/h (segundos)	-.196	.102	3.651	1	.056	.822	.673	1.005
	[Número de cilindros=3]	-36.916	6351.556	.000	1	.995	9.282E-17	.000	.[b]
	[Número de cilindros=4]	-18.154	571.747	.001	1	.975	1.305E-8	.000	.[b]
	[Número de cilindros=5]	-19.936	17542.032	.000	1	.999	2.197E-9	.000	.[b]
	[Número de cilindros=6]	-17.744	571.747	.001	1	.975	1.967E-8	.000	.[b]
	[Número de cilindros=8]	0[c]	.	.	0	.	.	.	.
Europa	Intersección	-4.007	975.389	.000	1	.997			
	Potencia (CV)	-.044	.021	4.178	1	.041	.957	.918	.998
	Peso total (kg)	.011	.003	15.872	1	.000	1.011	1.006	1.017
	Aceleración 0 a 100 km/h (segundos)	-.045	.102	.198	1	.657	.956	.783	1.167
	[Número de cilindros=3]	-20.400	.000	.	1	.	1.381E-9	1.381E-9	1.381E-9
	[Número de cilindros=4]	-.424	975.385	.000	1	1.000	.654	.000	.[b]
	[Número de cilindros=5]	17.909	14899.566	.000	1	.999	59936993,15	.000	.[b]
	[Número de cilindros=6]	-2.385	975.384	.000	1	.998	.092	.000	.[b]
	[Número de cilindros=8]	0[c]	.	.	0	.	.	.	.

La tabla anterior muestra para cada variable y para cada nivel de la variable categórica independiente el valor del coeficiente estimado (β), su error típico, el estadístico de Wald, los p-valores que miden el nivel de significancia de cada coeficiente en el modelo, la razón de las ventajas estimadas y el intervalo de confianza al 95% para las ventajas estimadas, se muestra en la columna sig que las variables peso potencia y aceleración tienen una fuerte significativa al momento de realizar una predicción de origen geográfico. Evaluación del ajuste del modelo.

3.2.3. Evaluación del ajuste del modelo

Para evaluar la bondad de ajuste se han calculado diferentes estadísticos (los cuales tratar de ser una especie de equivalentes al R^2 que conocemos), los cuales se muestran en la siguiente tabla.

Pseudo R cuadrado

Cox y Snell	.484
Nagelkerke	.575
McFadden	.358

Los valores de los Pseudo R^2 son aceptables, y muestran que los datos se ajustan bien al modelo predictivo, por ejemplo el calculo del Pseudo R^{2} de Mcfadden se calcula como la razón de máxima verosimilitud entre la deviance con la constante y eso ayuda a que se verifique si las variables explicativas realmente tienen influencia al momento de predecir el origen geográfico de un automóvil. Ya que, si la deviance calculada con todos los coeficientes es pequeña, entonces esa razón tenderá a 1 y significará que los datos se ajustan bien al modelo.

3.2.4. Capacidad discriminante del modelo

Ahora se muestra otro muy buen indicador para evaluar la capacidad predictiva de nuestro modelo de elección múltiple, en la siguiente tabla se tiene la matriz de confución, la cual es una tabla de continegencia de los origenes reales y los origenes predichos por el modelo de regresion logistica multinomial.

Clasificación

	Pronosticado			
Observado	EE.UU.	Europa	Japón	Porcentaje correcto
EE.UU.	219	8	22	88.0%
Europa	27	15	29	21.1%
Japón	23	3	53	67.1%
Porcentaje global	67.4%	6.5%	26.1%	71.9%

Como podemos observar en la tabla de clasificación, el poder clasificativo del modelo de elección discreta es muy bueno, debido a que ha clasificado de manera correcta a prácticamente el 72% de los mas de 400 automóviles que se tenían en la muestra para realizar la estimación de este modelo, y la mayor satisfacción que se puede tener es que este poder clasificativo se proyecte a estimaciones futuras que se realicen para determinar el origen geográfico de uno o un conjunto de automóviles.