Tarea1_CesarArancibia_FranciscoRios

September 26, 2022

Section 2: non linear models

[2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
import sklearn
import scipy
from scipy.stats import nbinom

%matplotlib inline
```

[4]:
```
junaeb = pd.read_csv(open('C:/Users/PC/Downloads/junaeb.csv'));
junaeb.dropna(inplace=True);
```

[5]:
```
from IPython.display import display
junaeb.reset_index(drop=True, inplace=True);
print(len(junaeb))
display(junaeb.head(5));
```

```
6379
```

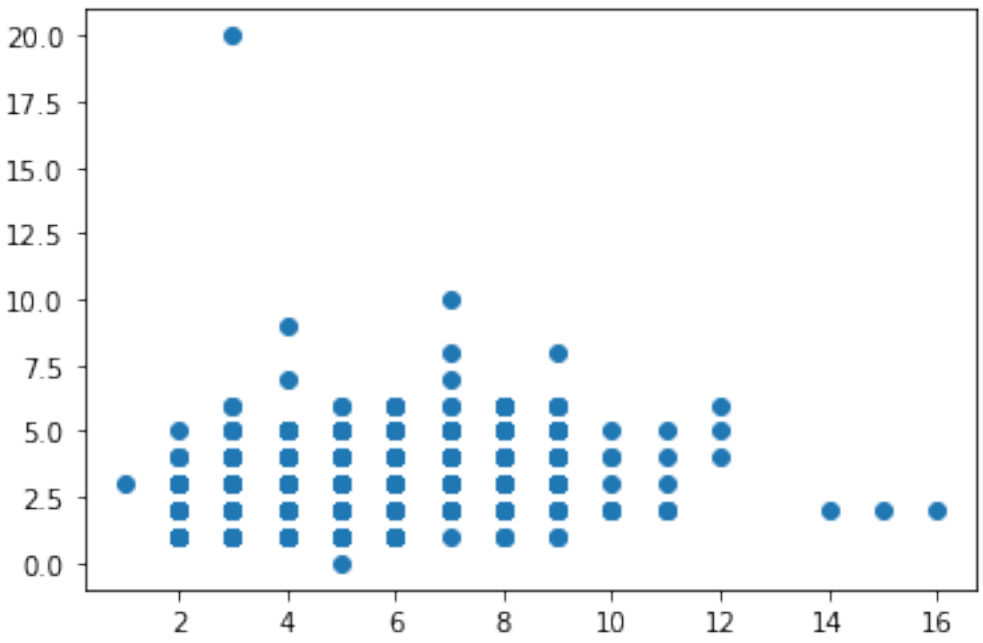
	vive_padre	vive_madre	n_personas	n_habitaciones	cercania_juegos	\
0	0	1	3.0	4.0	1.0	
1	0	1	5.0	3.0	1.0	
2	1	1	5.0	3.0	1.0	
3	1	1	4.0	2.0	1.0	
4	1	1	5.0	3.0	2.0	

	cercania_servicios	edad_primer_parto	area	educm	educp
0	1.0	25.0	1	0	0
1	1.0	23.0	1	13	13
2	1.0	19.0	1	12	17
3	1.0	27.0	1	6	13
4	1.0	20.0	1	13	16

[6]:
```
#junaeb.info();
junaeb.n_personas = junaeb.n_personas.astype(int);
junaeb.n_habitaciones = junaeb.n_habitaciones.astype(int);
junaeb.cercania_juegos = junaeb.cercania_juegos.astype(int);
junaeb.cercania_servicios = junaeb.cercania_servicios.astype(int);
junaeb.edad_primer_parto = junaeb.edad_primer_parto.astype(int);
junaeb.info();
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6379 entries, 0 to 6378
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   vive_padre          6379 non-null   int64
 1   vive_madre          6379 non-null   int64
 2   n_personas          6379 non-null   int32
 3   n_habitaciones      6379 non-null   int32
 4   cercania_juegos     6379 non-null   int32
 5   cercania_servicios  6379 non-null   int32
 6   edad_primer_parto   6379 non-null   int32
 7   area                6379 non-null   int64
 8   educm               6379 non-null   int64
 9   educp               6379 non-null   int64
dtypes: int32(5), int64(5)
memory usage: 373.9 KB
```

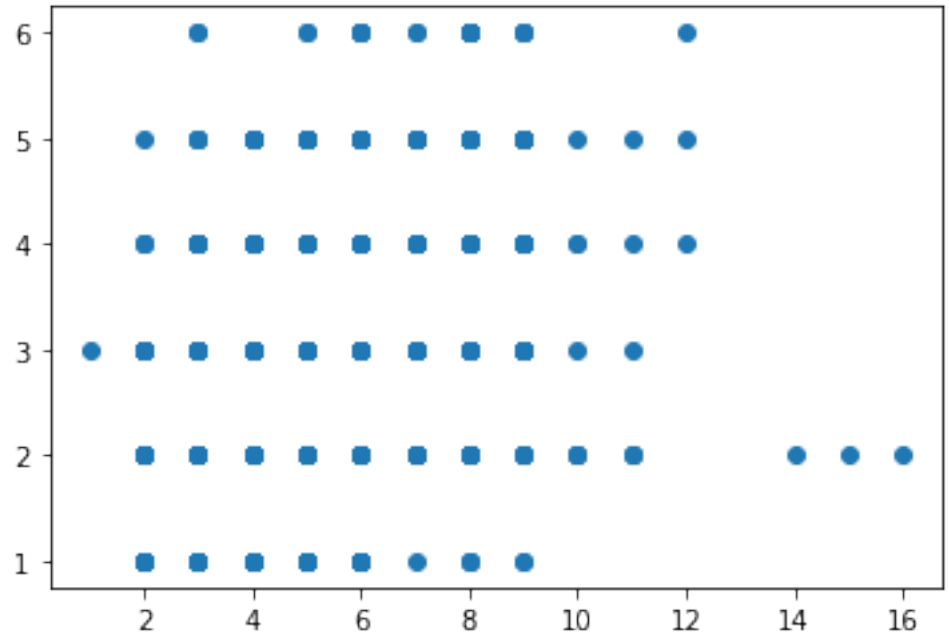
[7]:
```
plt.scatter(junaeb['n_personas'],junaeb['n_habitaciones']);
```



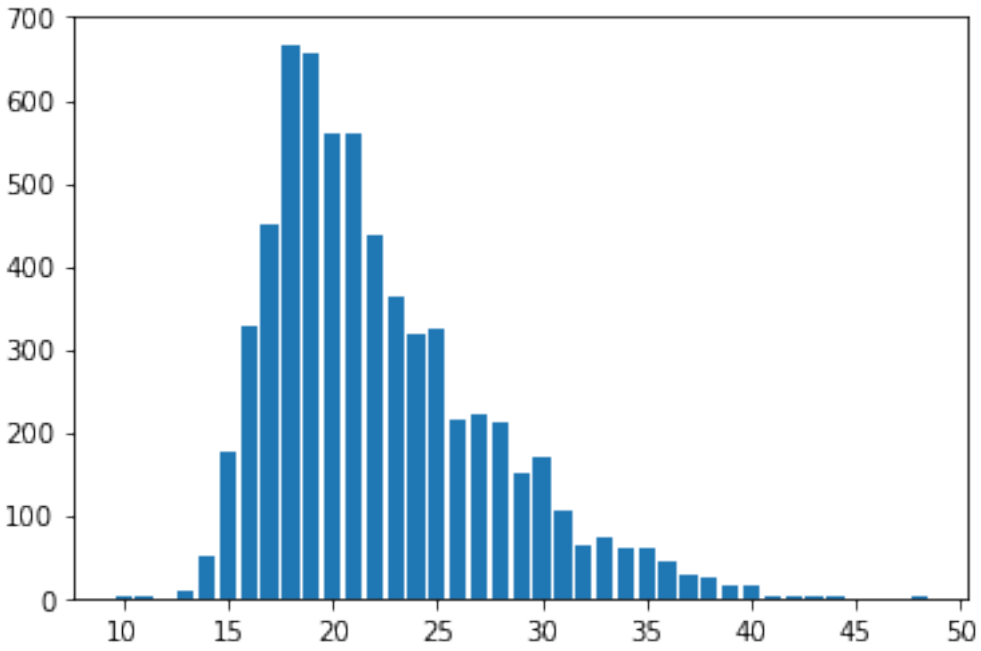
[8]:
```
#Del grafico se extrae que hay datos atipicas;
#Se decide eliminar datos atipicos: 20, 10, 9, 8 y 7 y 0 habitaciones
junaeb = junaeb[(junaeb["n_habitaciones"] < 7) & (junaeb["n_habitaciones"] >
↪0)];
print(len(junaeb));
```

```
6371
```

[9]:
```
plt.scatter(junaeb['n_personas'],junaeb['n_habitaciones']);
```



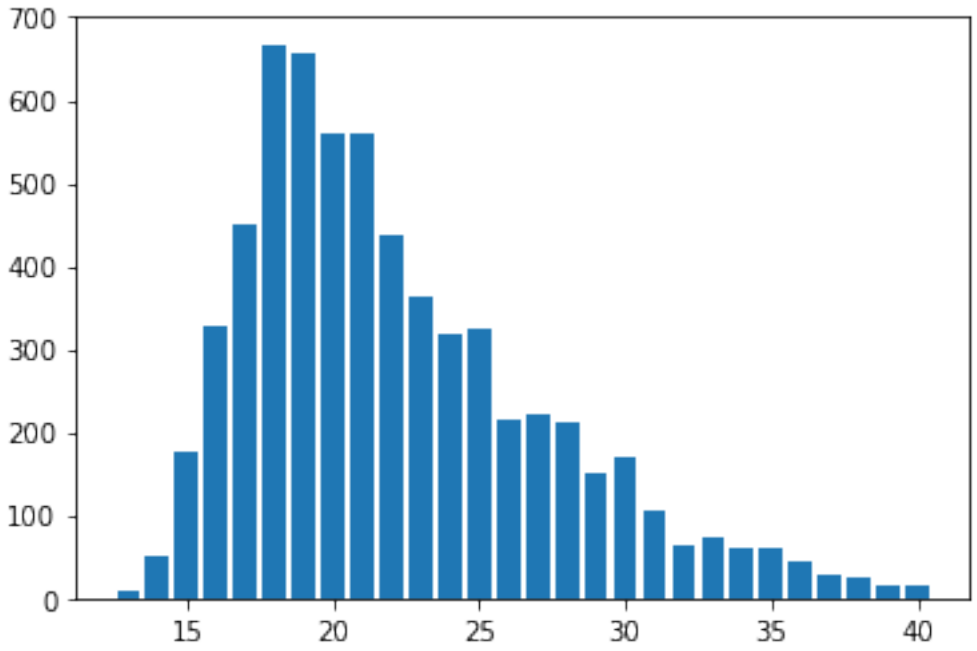
[10]:
```
graf = junaeb["edad_primer_parto"].value_counts().to_frame().reset_index();
graf.columns = ["edad_primer_parto", "rep"];
plt.bar(graf.edad_primer_parto, graf.rep);
plt.show();
```



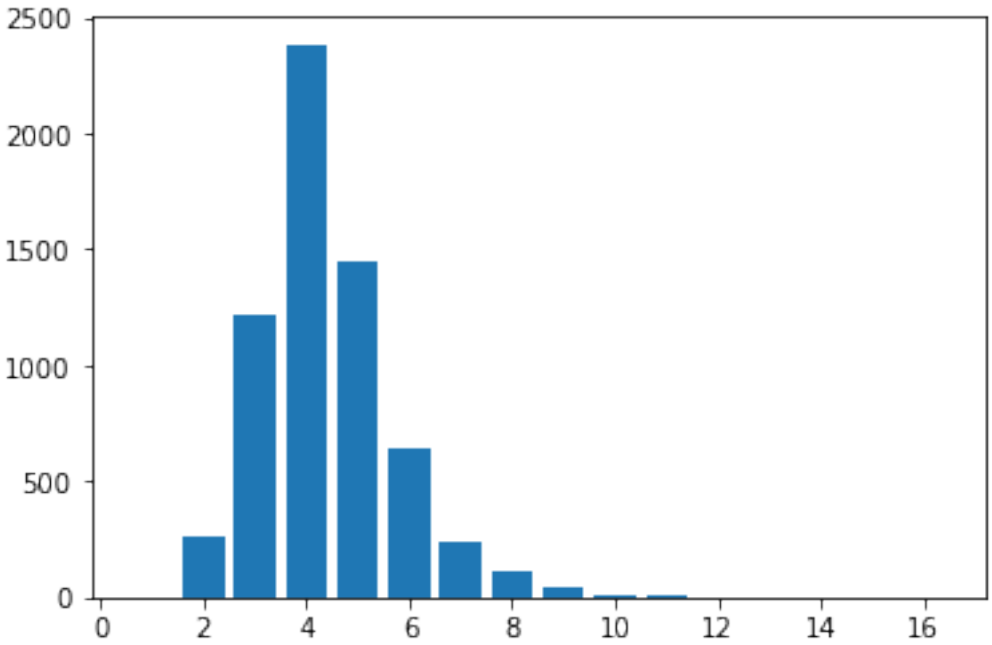
[11]:
```
#Se descubren datos atipicos
#Edad 10, 11 y mayores que 41 se van
junaeb = junaeb[(junaeb["edad_primer_parto"] > 12) &␣
 ↪(junaeb["edad_primer_parto"] <= 40)];
print(len(junaeb));

graf = junaeb["edad_primer_parto"].value_counts().to_frame().reset_index();
graf.columns = ["edad_primer_parto", "rep"];
plt.bar(graf.edad_primer_parto, graf.rep);
plt.show();
```

6359



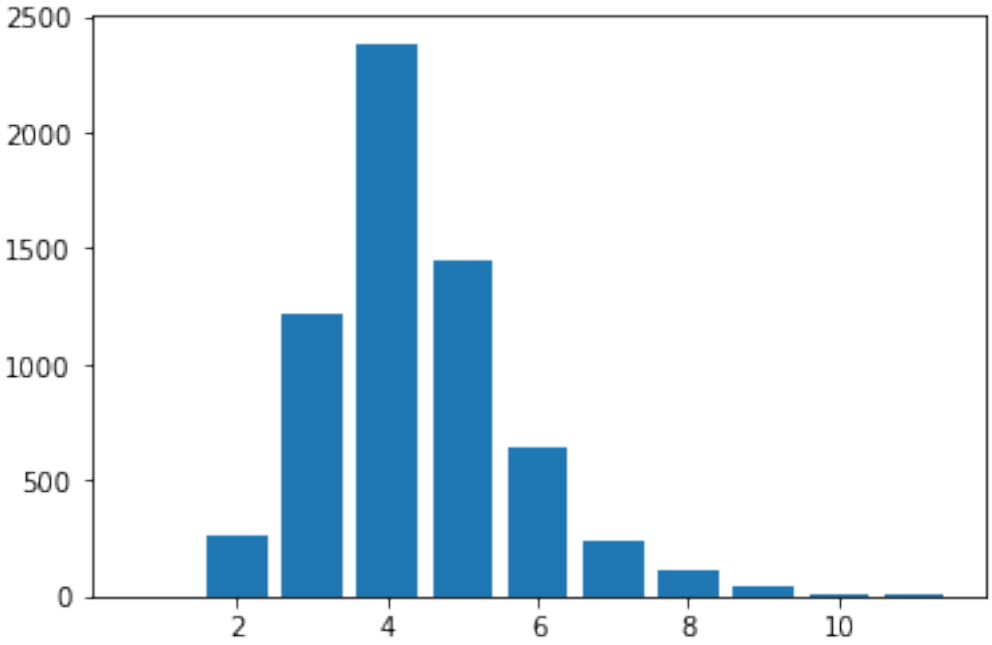
[12]:
```
graf = junaeb["n_personas"].value_counts().to_frame().reset_index();
graf.columns = ["n_personas", "rep"];
plt.bar(graf.n_personas, graf.rep);
plt.show();
```



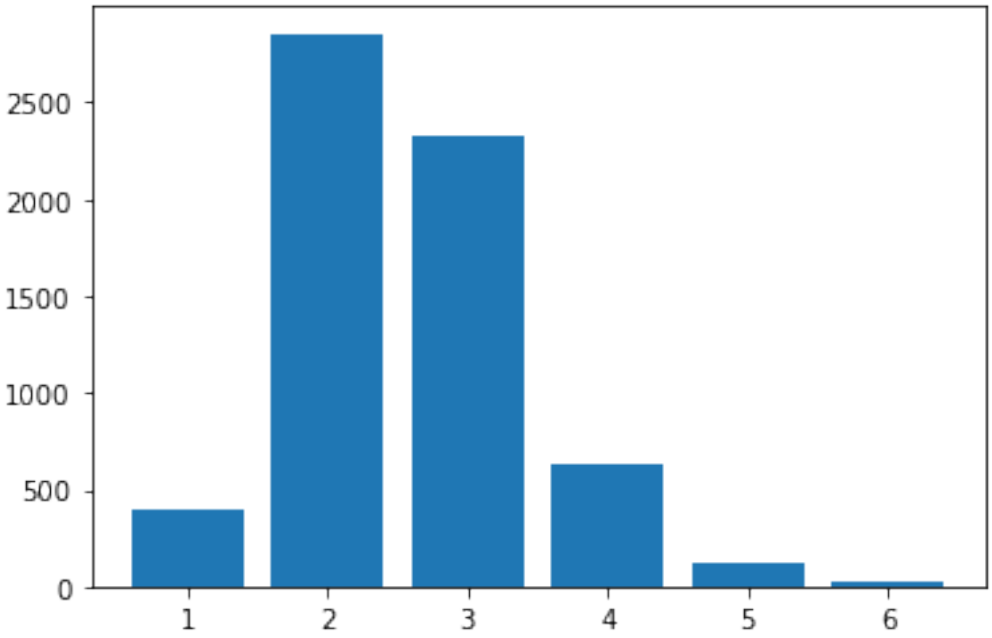
[13]:
```
#Se descubren datos atipicos
#No de personas 12, 14, 15 y 16 son atipicos
junaeb = junaeb[(junaeb["n_personas"] < 12)];
print(len(junaeb));

graf = junaeb["n_personas"].value_counts().to_frame().reset_index();
graf.columns = ["n_personas", "rep"];
plt.bar(graf.n_personas, graf.rep);
plt.show();
```

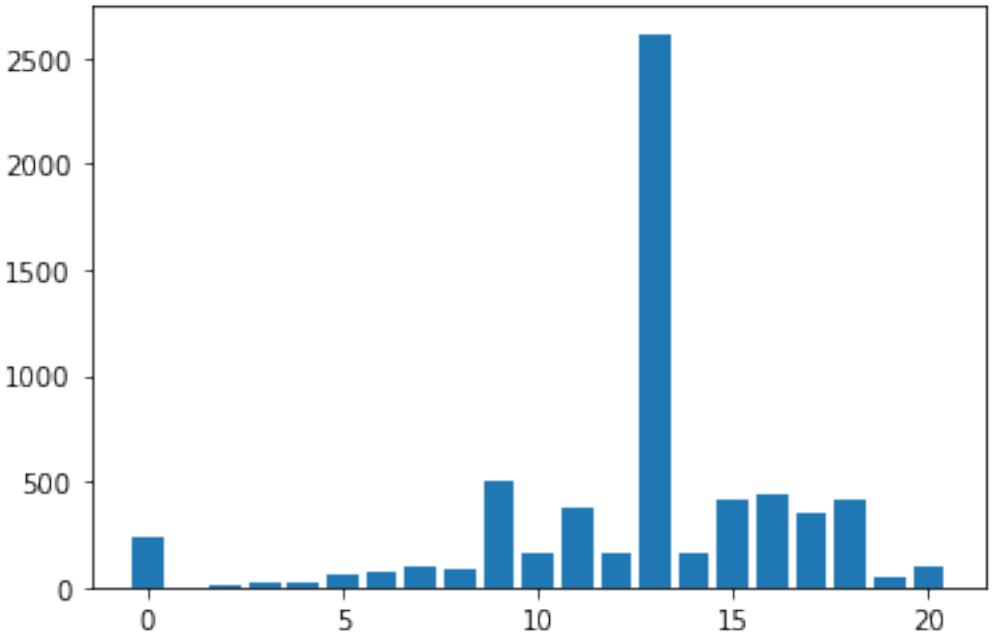
```
6353
```



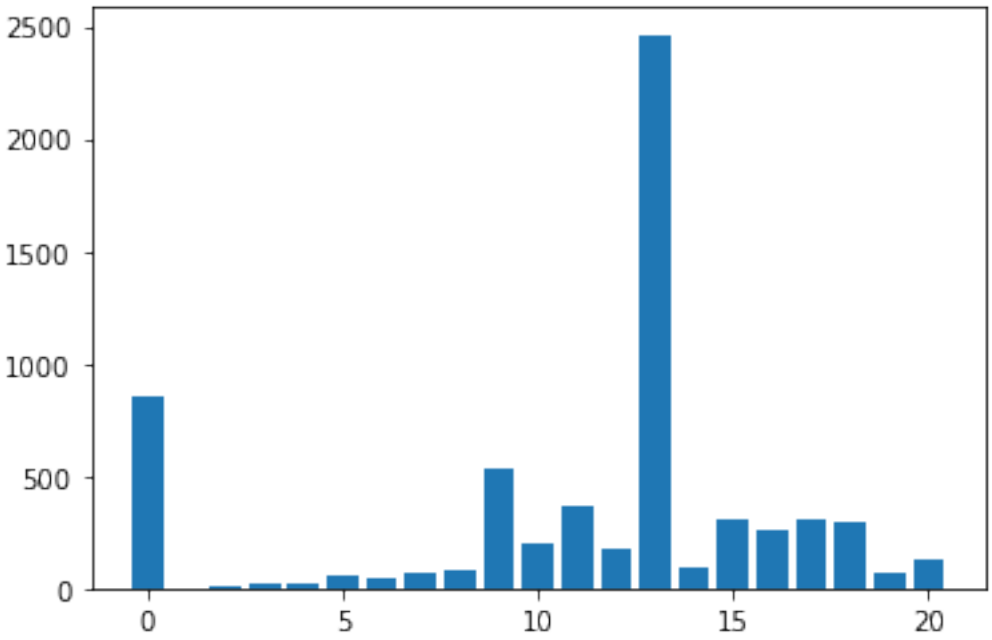
[14]:
```
graf = junaeb["n_habitaciones"].value_counts().to_frame().reset_index();
graf.columns = ["n_habitaciones", "rep"];
plt.bar(graf.n_habitaciones, graf.rep);
plt.show();
```



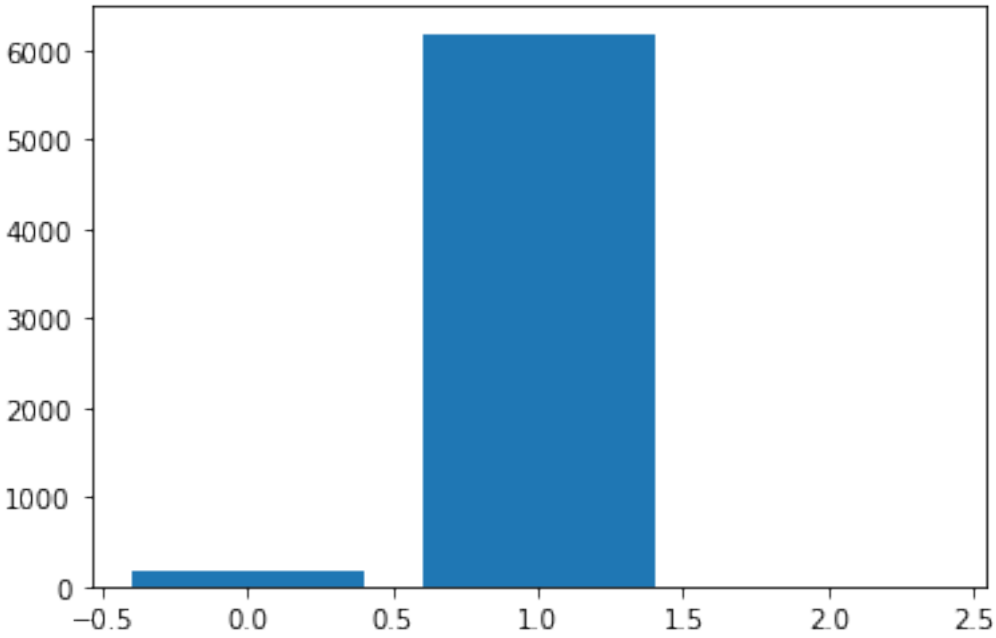
[15]:
```
graf = junaeb["educm"].value_counts().to_frame().reset_index();
graf.columns = ["educm", "rep"];
plt.bar(graf.educm, graf.rep);
plt.show();
```



[16]:
```
graf = junaeb["educp"].value_counts().to_frame().reset_index();
graf.columns = ["educp", "rep"];
plt.bar(graf.educp, graf.rep);
plt.show();
```



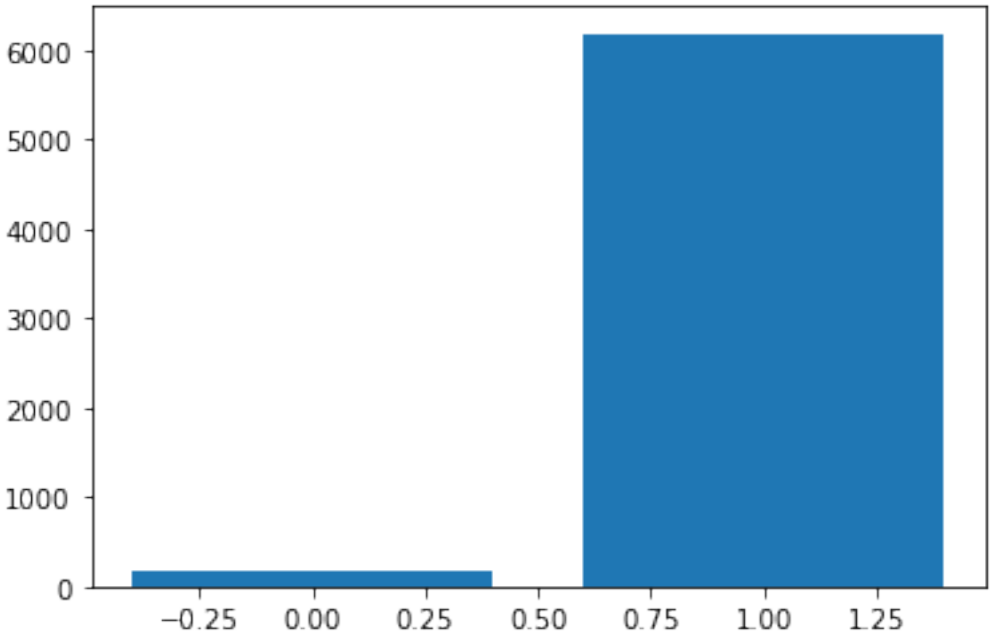
[18]:
```
graf = junaeb["vive_madre"].value_counts().to_frame().reset_index();
graf.columns = ["vive_madre", "rep"];
plt.bar(graf.vive_madre, graf.rep);
plt.show();
```



[19]:
```
#Se descubren datos atipicos
#Vive madre 2 es atipico
junaeb = junaeb[(junaeb["vive_madre"] < 2)];
print(len(junaeb));

graf = junaeb["vive_madre"].value_counts().to_frame().reset_index();
graf.columns = ["vive_madre", "rep"];
plt.bar(graf.vive_madre, graf.rep);
plt.show();
```

6348

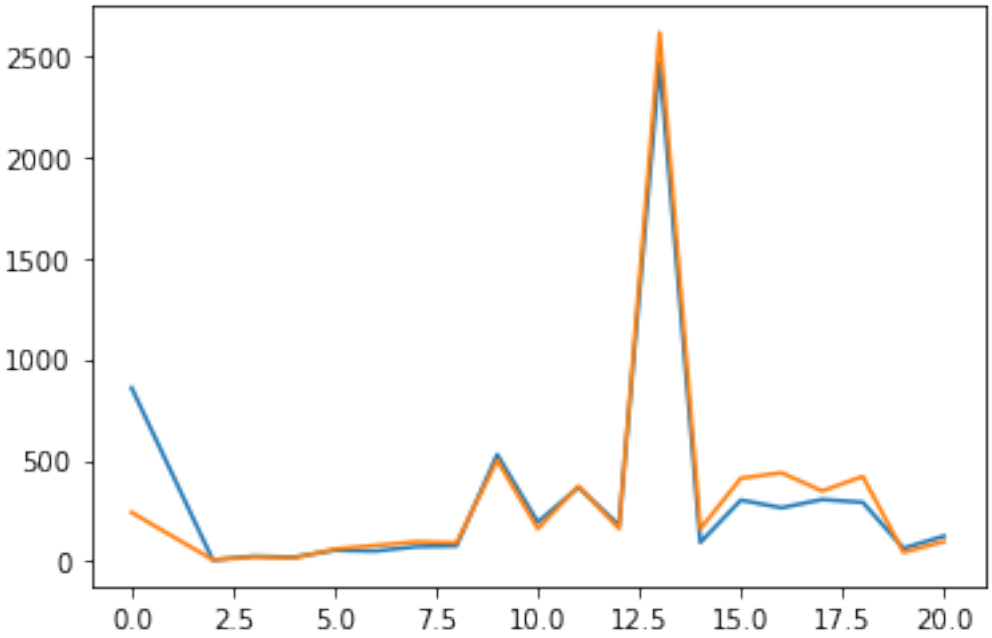


[20]:
```
#Grafico comparativo entre educp y educm

graf1 = junaeb["educp"].value_counts().to_frame().reset_index();
graf1.columns = ["educp", "rep"];
graf1 = graf1.sort_values(by=["educp"]);
plt.plot(graf1.educp, graf1.rep);

graf1 = junaeb["educm"].value_counts().to_frame().reset_index();
graf1.columns = ["educm", "rep"];
graf1 = graf1.sort_values(by=["educm"]);
plt.plot(graf1.educm, graf1.rep);

plt.show();
```



0.0.1 MCO

[21]:
```
y=junaeb['vive_padre']
X=junaeb[['vive_madre','n_personas','n_habitaciones','cercania_juegos','cercania_servicios','e
          'area','educm','educp']];
X=sm.add_constant(X);
model = sm.OLS(y, X);
results = model.fit();
print(results.summary());
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             vive_padre   R-squared:                       0.171
Model:                            OLS   Adj. R-squared:                  0.169
Method:                 Least Squares   F-statistic:                     144.9
Date:                Thu, 15 Sep 2022   Prob (F-statistic):          9.81e-250
Time:                        16:25:29   Log-Likelihood:                -3543.0
No. Observations:                6348   AIC:                             7106.
Df Residuals:                    6338   BIC:                             7174.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
======
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
```

```
------
const                  0.1150      0.054      2.146      0.032       0.010
0.220
vive_madre             0.1163      0.034      3.418      0.001       0.050
0.183
n_personas             0.0597      0.005     12.679      0.000       0.051
0.069
n_habitaciones        -0.0481      0.007     -6.758      0.000      -0.062
-0.034
cercania_juegos       -0.0109      0.013     -0.829      0.407      -0.037
0.015
cercania_servicios     0.0139      0.014      0.982      0.326      -0.014
0.042
edad_primer_parto      0.0099      0.001      9.174      0.000       0.008
0.012
area                  -0.0801      0.018     -4.364      0.000      -0.116
-0.044
educm                 -0.0161      0.001    -11.048      0.000      -0.019
-0.013
educp                  0.0334      0.001     31.128      0.000       0.031
0.035
==============================================================================
Omnibus:                      766.992   Durbin-Watson:                   1.983
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              817.535
Skew:                          -0.826   Prob(JB):                    2.98e-178
Kurtosis:                       2.397   Cond. No.                        316.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

0.0.2 Probit

[22]:
```
model = sm.Probit(y, X);
probit_model = model.fit();
print(probit_model.summary());

mfx = probit_model.get_margeff();
print(mfx.summary());
```

```
Optimization terminated successfully.
         Current function value: 0.536997
         Iterations 5
                          Probit Regression Results
==============================================================================
Dep. Variable:             vive_padre   No. Observations:                 6348
Model:                         Probit   Df Residuals:                     6338
```

```
Method:                           MLE   Df Model:                            9
Date:                Thu, 15 Sep 2022   Pseudo R-squ.:                  0.1375
Time:                        16:25:43   Log-Likelihood:                -3408.9
converged:                       True   LL-Null:                       -3952.3
Covariance Type:            nonrobust   LLR p-value:                3.260e-228
==============================================================================
======
                         coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
------
const                 -1.2316      0.176     -7.018      0.000      -1.576
-0.888
vive_madre             0.3247      0.109      2.983      0.003       0.111
0.538
n_personas             0.1892      0.016     12.081      0.000       0.159
0.220
n_habitaciones        -0.1503      0.023     -6.454      0.000      -0.196
-0.105
cercania_juegos       -0.0342      0.043     -0.804      0.421      -0.118
0.049
cercania_servicios     0.0381      0.046      0.821      0.412      -0.053
0.129
edad_primer_parto      0.0352      0.004      9.608      0.000       0.028
0.042
area                  -0.2444      0.061     -3.998      0.000      -0.364
-0.125
educm                 -0.0518      0.005    -10.412      0.000      -0.062
-0.042
educp                  0.0976      0.004     26.931      0.000       0.091
0.105
==============================================================================
======
       Probit Marginal Effects
=====================================
Dep. Variable:             vive_padre
Method:                          dydx
At:                           overall
==============================================================================
======
                        dy/dx    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
------
vive_madre             0.0992      0.033      2.988      0.003       0.034
0.164
n_personas             0.0578      0.005     12.429      0.000       0.049
0.067
```

```
n_habitaciones        -0.0459      0.007     -6.511      0.000      -0.060
-0.032
cercania_juegos       -0.0104      0.013     -0.804      0.421      -0.036
0.015
cercania_servicios     0.0116      0.014      0.821      0.411      -0.016
0.039
edad_primer_parto      0.0107      0.001      9.775      0.000       0.009
0.013
area                  -0.0747      0.019     -4.011      0.000      -0.111
-0.038
educm                 -0.0158      0.001    -10.606      0.000      -0.019
-0.013
educp                  0.0298      0.001     32.269      0.000       0.028
0.032
==============================================================================
======
```

0.0.3 Logit

[23]:
```
model = sm.Logit(y, X);
logit_model = model.fit();
print(logit_model.summary());

mfx = logit_model.get_margeff();
print(mfx.summary());
```

```
Optimization terminated successfully.
         Current function value: 0.534681
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:             vive_padre   No. Observations:                 6348
Model:                          Logit   Df Residuals:                     6338
Method:                           MLE   Df Model:                            9
Date:                Thu, 15 Sep 2022   Pseudo R-squ.:                  0.1412
Time:                        16:25:50   Log-Likelihood:                -3394.2
converged:                       True   LL-Null:                       -3952.3
Covariance Type:            nonrobust   LLR p-value:                1.474e-234
==============================================================================
======
                         coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
------
const                 -2.1518      0.300     -7.166      0.000      -2.740
-1.563
vive_madre             0.5865      0.182      3.226      0.001       0.230
0.943
```

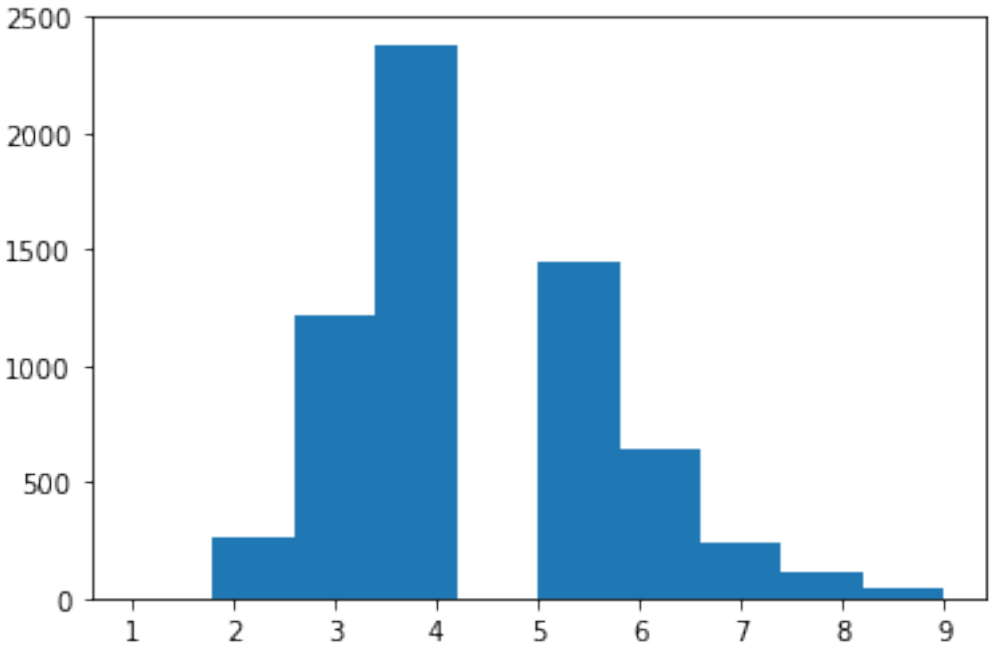
```
n_personas              0.3331      0.028     11.789      0.000       0.278
0.389
n_habitaciones         -0.2551      0.040     -6.315      0.000      -0.334
-0.176
cercania_juegos        -0.0596      0.072     -0.822      0.411      -0.201
0.082
cercania_servicios      0.0667      0.079      0.848      0.396      -0.087
0.221
edad_primer_parto       0.0621      0.006      9.682      0.000       0.050
0.075
area                   -0.4125      0.105     -3.928      0.000      -0.618
-0.207
educm                  -0.0961      0.009    -10.622      0.000      -0.114
-0.078
educp                   0.1674      0.006     26.061      0.000       0.155
0.180
==============================================================================
======
        Logit Marginal Effects
=====================================
Dep. Variable:             vive_padre
Method:                          dydx
At:                           overall
==============================================================================
======
                        dy/dx    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
------
vive_madre              0.1042      0.032      3.236      0.001       0.041
0.167
n_personas              0.0592      0.005     12.199      0.000       0.050
0.069
n_habitaciones         -0.0453      0.007     -6.377      0.000      -0.059
-0.031
cercania_juegos        -0.0106      0.013     -0.822      0.411      -0.036
0.015
cercania_servicios      0.0119      0.014      0.849      0.396      -0.016
0.039
edad_primer_parto       0.0110      0.001      9.888      0.000       0.009
0.013
area                   -0.0733      0.019     -3.943      0.000      -0.110
-0.037
educm                  -0.0171      0.002    -10.937      0.000      -0.020
-0.014
educp                   0.0297      0.001     32.710      0.000       0.028
0.032
==============================================================================
```

======

0.0.4 Poisson

[24]:
```
subset=junaeb.loc[junaeb['n_personas']<10];
y=subset['n_personas'];
X=subset[['vive_padre','vive_madre','edad_primer_parto','n_habitaciones','cercania_juegos','ce
          'area','educm','educp']];
plt.hist(subset.n_personas);
display(subset.n_personas.head());
```

```
0    3
1    5
2    5
3    4
4    5
Name: n_personas, dtype: int32
```



[25]:
```
poisson=sm.GLM(y,X,family=sm.families.Poisson()).fit();
print(poisson.summary());
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              n_personas   No. Observations:                 6335
Model:                             GLM   Df Residuals:                     6326
Model Family:                  Poisson   Df Model:                            8
```

```
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -11485.
Date:                Thu, 15 Sep 2022   Deviance:                       1987.5
Time:                        16:25:59   Pearson chi2:                 2.15e+03
No. Iterations:                     5
Covariance Type:            nonrobust
==============================================================================
======
                         coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
------
vive_padre             0.1194      0.014      8.338      0.000       0.091
0.147
vive_madre             0.5907      0.036     16.303      0.000       0.520
0.662
edad_primer_parto     -0.0036      0.001     -3.149      0.002      -0.006
-0.001
n_habitaciones         0.1995      0.006     31.137      0.000       0.187
0.212
cercania_juegos        0.0855      0.014      6.145      0.000       0.058
0.113
cercania_servicios     0.0754      0.015      5.051      0.000       0.046
0.105
area                   0.1487      0.021      7.162      0.000       0.108
0.189
educm                  0.0019      0.002      1.168      0.243      -0.001
0.005
educp                  0.0004      0.001      0.311      0.756      -0.002
0.003
==============================================================================
======
```

[26]:
```
print("fitted lambda")
print(poisson.mu)
```

```
fitted lambda
[4.99290279 4.24705432 4.85370002 … 3.26671171 3.83571846 4.16894752]
```

0.0.5 Binomial Negativo

[27]:
```
negbin=sm.GLM(y,X,family=sm.families.NegativeBinomial()).fit();
print(negbin.summary());
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:             n_personas   No. Observations:                 6335
Model:                            GLM   Df Residuals:                     6326
```

```
Model Family:        NegativeBinomial   Df Model:                            8
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -16328.
Date:                Thu, 15 Sep 2022   Deviance:                       379.15
Time:                        16:26:10   Pearson chi2:                     422.
No. Iterations:                     8
Covariance Type:            nonrobust
==============================================================================
======
                         coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
------
vive_padre             0.1277      0.033      3.909      0.000       0.064
0.192
vive_madre             0.4759      0.075      6.303      0.000       0.328
0.624
edad_primer_parto     -0.0025      0.003     -0.942      0.346      -0.008
0.003
n_habitaciones         0.2105      0.015     13.719      0.000       0.180
0.241
cercania_juegos        0.1021      0.033      3.104      0.002       0.038
0.167
cercania_servicios     0.0998      0.036      2.809      0.005       0.030
0.169
area                   0.1616      0.046      3.513      0.000       0.071
0.252
educm                  0.0022      0.004      0.564      0.573      -0.005
0.010
educp                  0.0005      0.003      0.161      0.872      -0.005
0.006
==============================================================================
======
```

[28]:
```
print("fitted lambda")
print(negbin.mu)
```

```
fitted lambda
[5.05206753 4.25726034 4.88381687 … 3.3016049  3.86177459 4.21715143]
```

0.0.6 Test de sobredispersión

A simple test for overdispersion can be determined with the results of the Poisson model, using the ratio of Pearson chi2 / Df Residuals. A value larger than 1 indicates overdispersion. In the case above (6), data suggets overdispersion.

The Negative Binomial model estimated above is using a value of θ (or $\alpha = 1/\theta$) equal to 1. In order to determine the appropiate value of α, you can estimate a simple regression using the output of the Poisson model:

1. Construct the following variable aux=$[(y-\lambda)^2-\lambda]/\lambda$
2. Regress the variable aux with λ as the only explanatory variable (no constant)
3. The estimated value is an appropiate guess for $\alpha = 1/\theta$

In the model of the previous section, just use the options on sm.families.NegativeBinomial, in order to manually enter the value of alpha. See example below.

[29]:
```
aux=((y-poisson.mu)**2-poisson.mu)/poisson.mu
auxr=sm.OLS(aux,poisson.mu).fit()
print(auxr.params)
print(auxr.summary())
```

```
x1   -0.150348
dtype: float64
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:             n_personas   R-squared (uncentered):
0.457
Model:                            OLS   Adj. R-squared (uncentered):
0.457
Method:                 Least Squares   F-statistic:
5323.
Date:                Thu, 15 Sep 2022   Prob (F-statistic):
0.00
Time:                        16:26:19   Log-Likelihood:
-6949.7
No. Observations:                6335   AIC:
1.390e+04
Df Residuals:                    6334   BIC:
1.391e+04
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
x1            -0.1503      0.002    -72.959      0.000      -0.154      -0.146
==============================================================================
Omnibus:                     7450.400   Durbin-Watson:                   1.971
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1037972.793
Skew:                           6.143   Prob(JB):                         0.00
Kurtosis:                      64.493   Cond. No.                         1.00
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

[25]:
```
negbin=sm.GLM(y,X,family=sm.families.NegativeBinomial(alpha=(0.15))).fit();
print(negbin.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:             n_personas   No. Observations:                 6340
Model:                            GLM   Df Residuals:                     6331
Model Family:        NegativeBinomial   Df Model:                            8
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -12712.
Date:                Mon, 12 Sep 2022   Deviance:                       1205.5
Time:                        21:18:07   Pearson chi2:                 1.33e+03
No. Iterations:                     6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
vive_padre     0.1241      0.018      6.766      0.000       0.088       0.160
vive_madre     0.5259      0.044     12.066      0.000       0.440       0.611
edad_primer_parto    -0.0030      0.001     -2.070      0.038      -0.006      -0.000
n_habitaciones     0.2058      0.008     24.512      0.000       0.189       0.222
cercania_juegos     0.0946      0.018      5.233      0.000       0.059       0.130
cercania_servicios     0.0871      0.019      4.482      0.000       0.049       0.125
area           0.1574      0.026      6.019      0.000       0.106       0.209
educm          0.0022      0.002      1.012      0.312      -0.002       0.006
educp          0.0004      0.002      0.258      0.797      -0.003       0.004
==============================================================================
```

Tarea 1

Instrucciones

Los resultados de los ejericicios propuestos se deben entregar como un notebook por correo el dia 14/9 hasta las 21:00. Además, es importante considerar que para que la revisión se pueda llevar a cabo, el código debe poder ejecutarse en cualquier computadora.

Las variables tienen la siguiente descripcion:

- vive_padre: si el padre vive en el hogar
- vive_madre: si la madre vive en el hogar
- n_personas: numero de integrantes del hogar
- n_habitaciones: numero de cuartos en el hogar
- cercania_juegos: hay juegos infantiles cerca de la vivienda (1=no, 2=si, 4=no sabe)
- cercania_servicios: hay serivicios de salud cerca de la vivienda (1=no, 2=si, 4=no sabe)
- edad_primer_parto: edad de la madre en su primer parto
- area: urbana=1, rural=0
- educm: años de escolaridad de la madre
- educp: años de escolaridad del padre

Preguntas:

1. Cargar la base de datos *junaeb.csv* en el ambiente. Identifique los tipos de datos que se encuentran en la base, realice estadísticas descriptivas sobre las variables importantes (Hint: Revisar la distribuciones, datos faltantes, outliers, etc.) y limpie las variables cuando sea necesario.

R: Al momento de analizar las variables presentadas en el archivo .csv, se notaron varias distribuciones beta y se detectaron los outliers. Estos últimos fueron eliminados, tal y como se muestra en el notebook, debido a que son datos atípicos que pueden afectar los resultados de los modelos.

2. Ejecute un modelo de probabilidad lineal (*MCO*) que permita explicar la probabilidad de que los padres se encuentren viviendo en el hogar (*vive_padre*). Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: Se utiliza un modelo MCO para ajustar el modelo. La obtención de un R ajustado igual a 17% nos deja ver que muy probablemente (sin contar la comparación con informes similares), las variables x no explican completamente a la variable dependiente. Esto era de suponer al utilizar una regresión lineal para estimar una variable binaria. Se extrae que solo dos de las variables utilizadas no son significativas: cercanía a centros de salud y cercanía a juegos. La variable significativa de la cantidad de habitaciones se relaciona negativamente con la estadía del padre en el hogar, siendo menos probable que éste viva ahí por cada pieza añadida a la variable en 0,0482 unidades. Para la aplicación de este modelo se supone un valor esperado del error = 0, caso que no se da en esta oportunidad.

3. Ejecute un modelo *probit* que permita explicar la probabilidad de que los padres se encuentren viviendo en el hogar (*vive_padre*). Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: Se supone una distribución del error normal y se aplica el modelo probit. Nuevamente las únicas variables no significativas en el modelo son la de cercanía a juegos y asistencias de salud, siendo la variable con mayor incidencia en vive_padre la variable explicativa vive_madre. En este caso, el padre tiene una probabilidad de 9,2% mayor de formar parte del hogar si es que la madre también vive en el hogar. Caso contrario al caso de las n habitaciones con el cual (en base al cambio marginal dy/dx) tiene un comportamiento inverso entre el aumento porcentual de las variables.

4. Ejecute un modelo *logit* que permita explicar la probabilidad de que los padres se encuentren viviendo en el hogar(*vive_padre*). Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: Al asumir que la distribución es logística se aumenta el valor de R ajustado a un 14%, obteniendo valores similares y descartando nuevamente solo las variables de cercanía. La función de maximización de verosimilitud representada por el Log verosimilitud dentro del panel entregado por statsmodel también sufre una variación positiva que indicaría una mejora en el uso de la regresión Logit. Se asume que la función se distribuye de forma logística y la variación del cambio marginal de las variables sobre la variable dependiente y es mínimo.

5. Comente los resultados obtenidos en 2, 3 y 4. ¿Cuáles y por qué existen las diferencias entre los resultados?. En su opinión, ¿Cuál sería el más adecuado para responder la pregunta de investigación y por qué? ¿Qué variables resultaron ser robustas a la especificación?

R: Es importante notar que gracias a lo estudiado se sabe que un modelo de regresión lineal no es recomendable para una variable dependiente binaria (cómo se aplicó en la pregunta 2) ya que toma valores fuera de los límites apropiados de estimación binaria 0 y 1. En ese sentido la opción más recomendable es la logit o probit. La logit en específico se recomienda por la asunción de distribución logística de los errores sin tener que asumir normalidad, aunque los resultados son similares. Los errores estándar y el cambio porcentual entre los pseudo-R2 favorable, junto al cambio del log verosimilitud nos indica que la aplicación del modelo logit logra la estimación más cercana de la probabilidad de que el resultado de la regresión se ajuste adecuadamente.

6. Ejecuta un modelo Poisson para explicar el número de personas que hay dentro de un hogar. (*n_personas*). Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: La función de distribución Poisson se utiliza para limitar el espacio muestral dentro de paredes que permitan analizar los resultados del experimento de mejor manera. En este caso los límites no son temporales sino físicos, al usar como variable de conteo la cantidad de habitaciones en el hogar, analizando los datos donde la cantidad de habitaciones era menor o igual a 6. Este valor es seleccionado al marcar el punto de inflexión donde comenzaba una cantidad de datos poco realistas y posibles outliers analíticos. El cambio en el Log likelihood entrega dudas sobre la forma de ajustar el modelo con el tipo de limitante, al tener un valor menor a -11000. Las variables expresan los cambios a Y a modo de coeficientes, que vendrían siendo el cambio porcentual en el número de eventos alrededor de la media por cada variable. En este caso, podemos observar como la variable vive_madre aumenta a la variable Y en e**0,57 veces por cada unidad extra de la variable.

7. Determine sobre dispersión y posible valor óptimo de alpha para un modelo Binomial Negativa.

R: Al determinar la dispersión de los residuos de Poisson, se obtiene un valor negativo de alpha (-0.15), lo que indicaría un resultado de underdisperssion, asegurando así lo previsto anteriormente. Con tal de tener la necesidad de aplicar una binomial negativa que controle los datos utilizados en la distribución de Poisson sería necesario tener un valor de alpha mayor que 1 para asegurar la existencia de sobredispersión.

8. Usando la información anterior, ejecute un modelo Binomial Negativa para explicar el número de personas que hay dentro de un hogar. (*n_personas*). Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: Aunque en este ejercicio en específico no se necesitó de la aplicación de un modelo de distribución binomial negativo al no presentarse sobredispersión, se utilizó un valor de 1 con los datos con tal de visualizar los resultados. En general, el uso de la Binomial negativa se justifica cuando los residuos de la distribución Poisson no son lo suficientemente claros, junto con una diferencia en la media

versus la varianza de Y|x. A modo de analogía, la distribución Poisson coloca entre cuatro paredes el experimento para poder observar mejor, mientras que la Binomial negativa coloca un techo que limita la cantidad de experimentos posibles a los observados.

Este modelo arrojó resultados similares al modelo Poisson, siendo la diferencia más grande que la variable explicativa edad_primer_parto no es considerada significativa o robusta por la binomial negativa, pero sí por la Poisson. Otras variables no significativas fueron educm y educp, mientras que todas las demás variables consideradas en el modelo fueron significativas.

9. Comente los resultados obtenidos en 6, 7 y 8. ¿Cuáles y por qué existen las diferencias entre los resultados?. En su opinión, ¿Cuál sería el más adecuado para responder la pregunta de investigación y por qué? ¿Qué variables resultaron ser robustas a la especificación?

R: La comparación se da entre el modelo Poisson de la respuesta 6 y el modelo binomial negativo de la respuesta 8. En este caso, al analizar los resultados obtenidos para los coeficientes de ambas regresiones se descubre que no hay una gran diferencia; por ejemplo, para el caso de vive_padre el modelo Poisson arroja un coeficiente de 0.1194 mientras que el modelo binomial negativo entrega un coeficiente de 0.1277. Sin embargo, el modelo binomial negativo considera que la variable edad_primer_parto no es significativa, mientras que el modelo Poisson sí la considera. Las diferencias pueden estar explicadas por el hecho de que se utiliza un alpha igual a 1 por defecto, puesto que no puede ser modificado por el valor estimado debido a que es un valor negativo. Por la razón anterior, se propone que el modelo Poisson es el mejor candidato para resolver este problema, ya que al no existir overdisperssion no hay una verdadera necesidad de ocupar el modelo binomial negativo. Finalmente, todas las variables son significativas o robustas con excepción de educm y educp.