Maximum Likelihood Estimation and discrete choice models

Monday, September 23

Open in Colab
(https://colab.research.google.com/github/arnauddyevre/Python-for-Social-Scientists/blob/master/statistics_and_econometrics/MLE/MLE_and_discrete_choice.ipynb)

Content

1. Short intro to MLE

Maximum likelihood is an intuitive and popular method of statistical inference. It involves choosing a set of parameters such that, when fed to a Data Generating Process (DGP), they match the observed moments of the data. The main advantages of maximum likelihood estimation are:

- its **efficiency**: It achieves the **Cramer-Rao (https://en.wikipedia.org/wiki/Cram%C3%A9r%E2%80%93Rao_bound)** lower bound when $n \rightarrow \infty$
- its **consistency**: The ML estimator converges in probability to the true parameter $\hat{\theta}_{MLE} \xrightarrow{p} \theta$, and even almost surely under stricter regularity conditions
- its **versatility**: Any DGP can be estimated by MLE. It is more versatile than linear regression as it allows for more intricate relationships between dependent variables.

Its main drawback is that it requires us to assume a specific parametric distribution of the data. Moreover, the maximisation algorithm may not find the global maximum.

2. Coding the MLE estimator

2.1. The linear case

Let's create a random sample of size $N = 1,000$, generated by the following DGP:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

The normality of errors and their 0-correlation ensure they are independent, a property we will use when implementing the MLE method.

In [1]:

```
import numpy as np
import scipy as sp
import pandas as pd
import matplotlib as mp
%matplotlib nbagg
import matplotlib.pyplot as plt

import statsmodels as sm
```

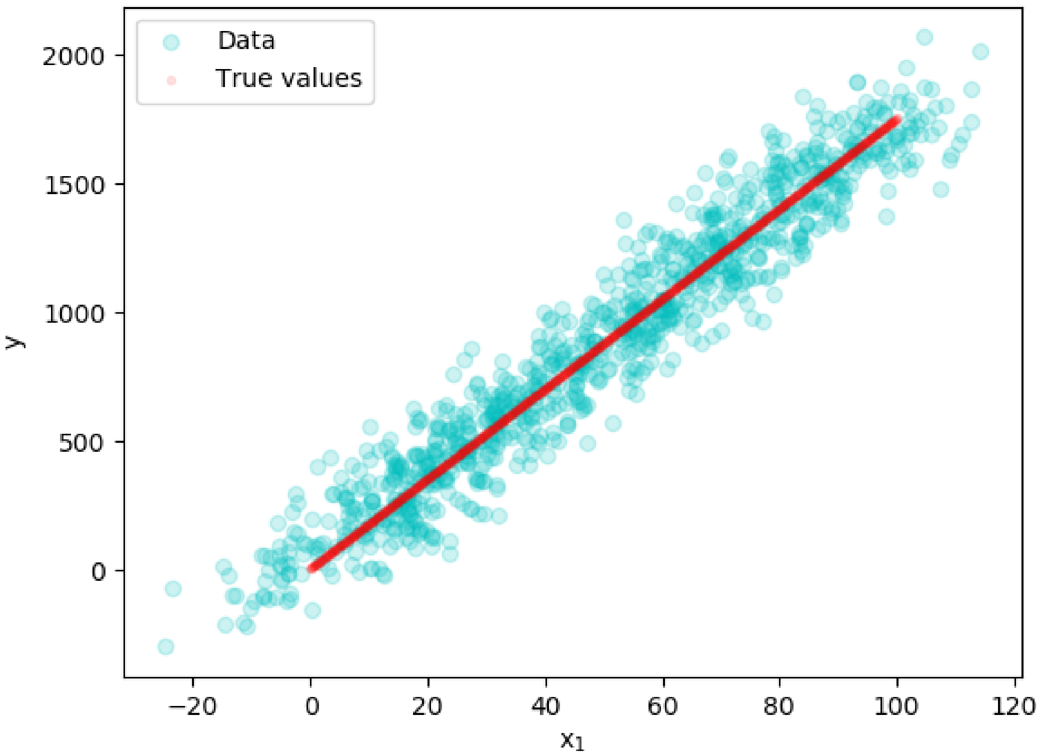
In [2]:

```
# True values of X's will be linear functions measured without noise
X1_true = np.linspace(0, 100, 1000)
X2_true = np.linspace(0, -500, 1000)
X3_true = np.linspace(0, 2000, 1000)

# We also add a vector of ones to the matrix of observables, this will allow us
 to include an intercept in the model
X0 = np.ones(1000)

X_true = np.array([X0,
              X1_true,
              X2_true,
              X3_true]).T    # Note the necessary transposition of the matrix

# We define observed covariates as the sum of the true X's + normal noise (can b
e interpreted as measurement error or simply as a way to avoid perfect multicoli
nearity)
np.random.seed(123)
X1 = X1_true + np.random.normal(loc=0.0, scale = 10, size=1000)
X2 = X2_true + np.random.normal(loc=0.0, scale = 50, size=1000)
X3 = X3_true + np.random.normal(loc=0.0, scale = 200, size=1000)

X = np.array([X0,
              X1,
              X2,
              X3]).T

# True parameters
β = np.array([5, 10, -.5, .25]).T
sigma = 100
epsilon = np.random.normal(loc=0.0, scale = sigma, size=1000)

# Measured and true values
y = X @ β + epsilon
y_true = X_true @ β
```

We get a simple linear, homoskedastic relastionship between the covariates and the observed y_i. See for instance $\mathbf{y}$ with respect to $\mathbf{x_1}$ below

In [3]:

```
plt.scatter(X1, y, c="c", alpha=0.2, label="Data")
plt.scatter(X1_true, y_true, c="r", alpha=0.1, marker='.', label = "True values"
)
plt.xlabel("$\mathrm{x_1}$")
plt.ylabel("$\mathrm{y}$")
plt.legend()
```



Out[3]:

```
<matplotlib.legend.Legend at 0x115785f28>
```

OLS is BLUE in this case. Now if we estimate the coefficients by OLS we get $\hat{\beta}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

In [4]:

```
# we used the dot() command last time for matrix multiplication, '@' does the sa
me thing for Python 3.5 and later versions
β_ols = np.linalg.inv((X.T @ X)) @ (X.T @ y)
β_ols
```

Out[4]:

```
array([11.83699744,  9.65608284, -0.57041114,  0.24929886])
```

Now under normality of errors, OLS and MLE estimates of the coefficients are asymptotically equal. Let's define an iterative procedure to estimate $\boldsymbol{\beta}$ through MLE and verify that the results are in line with OLS. When errors are normally distributed, we do not need to go through such lengths as the MLE estimator has a closed-form solution. But we simply show how the convergence algorithm works in this simple case.

Our true parameter values are

$$\boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 5 \\ 0.5 \\ 0.25 \end{bmatrix}, \sigma = 100$$

We define a likelihood function, based on N draws of the data

$$f\left(\mathbf{y}|\mathbf{X}; \boldsymbol{\beta}, \sigma^2\right) = \prod_{i=1}^{N} f\left(y_i|x_i, \boldsymbol{\beta}, \sigma^2\right)$$

where we assume that errors are normally i.i.d, so

$$f\left(y_i|x_i, \boldsymbol{\beta}, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{1}{2\sigma^2}\left(y_i - x_i'\boldsymbol{\beta}\right)^2}$$

We further define the likelihood function $\mathcal{L}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ which treats the parameters $\boldsymbol{\beta}$ and σ as random and the values $\mathbf{y}, \mathbf{X}$ as given. MLE consists in maximising the value of $\mathcal{L}(\boldsymbol{\beta}, \sigma|\mathbf{y}, \mathbf{X})$ by chosing $\boldsymbol{\beta}$ and σ optimally. It is easier to work with the log of the likelihood function and this does not affect the maximiser as any monotonically increasing transformation of a function has the same maximiser.

We thus maximise:

$$\begin{aligned} \ln \mathcal{L}\left(\boldsymbol{\beta}, \sigma|\mathbf{y}, \mathbf{X}\right) =& \ell\left(\boldsymbol{\beta}, \sigma|\mathbf{y}, \mathbf{X}\right) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - x_i'\beta\right)^2 \end{aligned}$$

And our solution is:

$$\left(\hat{\beta}_{MLE}, \hat{\sigma}_{MLE}\right) = \underset{\theta,\beta}{\operatorname{argmax}} \left\{ -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - x_i'\beta\right)^2 \right\}$$

2.2. Manual coding of the MLE problem

We now turn to coding the problem stated above. We aim to obtain MLE estimates based on the dataset created at the beginning of the Notebook.

We transform the problem into a minimisation one. Minimisation problems are more numerically stable. We will call a Scipy function for our problem: `scipy.optimize.minimize` .

The Scipy `minimize` function provides a vast collection of constrained and unconstrained minimizations algorithms for multivariate scalar functions. The default algorithm `minimize` resorts to when solving an unconstrained optimization problem (like ours) is the [Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (https://en.wikipedia.org/wiki/Broyden%E2%80%93Fletcher%E2%80%93Goldfarb%E2%80%93Shanno_alg](https://en.wikipedia.org/wiki/Broyden%E2%80%93Fletcher%E2%80%93Goldfarb%E2%80%93Shanno_alg) which is also used by MATLAB and R's optimisation routines. Another popular option is the [Nelder-Mead method (https://en.wikipedia.org/wiki/Nelder%E2%80%93Mead_method)](https://en.wikipedia.org/wiki/Nelder%E2%80%93Mead_method), whose robustness to many types of objective functions makes it a very versatile -but slow- algorithm. You can pass the algorithm you want use as an argument of `minimize` , based on your application.

We will simply use three arguments in our application of `minimize` :

- The function to be minimised, called the criterion (here $-\ell\left(\boldsymbol{\beta}, \sigma|\mathbf{y}, \mathbf{X}\right)$)

- An initial guess for the value of the β and σ parameters
- The data needed to be fed to the criterion

We first encode the criterion below. We need to be careful about the type of arguments taken by this function. Whether the parameters need to be entered as tuples, arrays, or lists is dictated by the how the `minimize` function works.

In [5]:

```
def negLogLNorm(params, *args):
    '''
    ------------------------------------------------------------------
    Calculate -log(likelihood) using data points "data" and parameters
    "params".
    The log(likelihood) is calculated based on the assumption that
    errors are normally distributed.
    ------------------------------------------------------------------

    INPUTS:
    params  =   numpy array with 2 elements: one k*1 vector of beta
                coefficients (including intercept), and one scalar for
                sigma
    data    =   numpy array with two elements: one N*1 vector of
                observations of the dependent variable, and one N*k
                matrix of observations of the dependent variables

    RETURNS:
    neg_log_lik_val =   The negative of the log likelihood, using the
                        assumption that errors are normally distributed
    ------------------------------------------------------------------
    '''

    # Fetching parameters and data (note: args and param are tuples)
    b0, b1, b2, b3, sigma = params
    β = np.array([b0, b1, b2, b3])    #Transforming the tuple of coefficients in
to an Numpy array to be used in matrix multiplication
    y, X = args[0], args[1]

    # Calculating bits of the log-likelihood
    E = y - X @ β                     # vector of errors
    SSE = np.square(E).sum()          # sum of squared errors
    N = len(y)                        #sample size

    # Define the log likelihood function
    log_likelihood = - (N/2)*np.log(2 * np.pi) - (N/2)*np.log(sigma**2) - (1/(2
* sigma**2))*SSE

    neg_log_likelihood = - log_likelihood

    return neg_log_likelihood
```

We can now use the `minimize` function. We import the `optimize` library under a convenient name, define some initial parameter values (our guesses), and define the data to be used.

In [6]:

```
import scipy.optimize as opt

β_0 = np.array([0, 0, 0, 0, 1])
data = (y, X)
MLE_results = opt.minimize(negLogLNorm, β_0, args=(data))
MLE_results
```

Out[6]:

```
      fun: 6011.944116143887
 hess_inv: array([[ 1.49266916e+01, -2.33223967e-03,  5.06954136e-0
3,
        -1.32404452e-02, -1.49125767e+01],
       [-2.33223967e-03,  3.55427102e-02, -2.00390865e-03,
        -2.19816527e-03,  4.52719406e-03],
       [ 5.06954136e-03, -2.00390865e-03,  1.15587576e-04,
         1.19462983e-04, -5.18854039e-03],
       [-1.32404452e-02, -2.19816527e-03,  1.19462983e-04,
         1.52618971e-04,  1.30873557e-02],
       [-1.49125767e+01,  4.52719406e-03, -5.18854039e-03,
         1.30873557e-02,  1.48990911e+01]])
      jac: array([ 6.10351562e-05,  2.44140625e-04, -1.15966797e-03,
5.12695312e-03,
        1.22070312e-04])
  message: 'Desired error not necessarily achieved due to precision
loss.'
     nfev: 687
      nit: 82
     njev: 98
   status: 2
  success: False
        x: array([11.83753269,  9.6560806 , -0.57041127,  0.2492985
8, 98.79104398])
```

`minimize` returns our estimated parameters (`x`), and much more, as an `OptimizeResult` object. We get the value of the negative log likelihood, Jacobian and Hessian matrices at the solution, whether the optimiser has converged (in our case it hasn't).

As anticipated, the estimated parameters are the same as those found via OLS. Our encoding of MLE has worked.

If you are interested in coding your own optimisation algorithm, the [Quant-Econ MLE lecture (https://lectures.quantecon.org/py/mle.html)](https://lectures.quantecon.org/py/mle.html) has a nice explanation and implementation of the [Newton-Raphson algorithm (https://en.wikipedia.org/wiki/Newton%27s_method)](https://en.wikipedia.org/wiki/Newton%27s_method).

2.3. Variance-covariance of $\hat{\theta}_{MLE}$

Note that we can easily obtain the variance-covariance matrix of our MLE estimator, as the inverse of the estimated Hessian is reported in the output above. Formally, the variance-covariance matrix of $\hat{\theta}_{MLE}$ is:

$$\text{var}(\theta) = \left(-E\left[\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial\theta\partial\theta'}\right]\right)^{-1}$$

We get it as follows:

In [7]:

```
var = MLE_results.hess_inv
se_alpha = np.sqrt(var[0, 0])
se_beta1 = np.sqrt(var[1, 1])
se_beta2 = np.sqrt(var[2, 2])
se_beta3 = np.sqrt(var[3, 3])
se_sigma = np.sqrt(var[4, 4])

print('SE(alpha) = ', se_alpha,
      '\nSE(beta_1) =', se_beta1,
      '\nSE(beta_2) =', se_beta2,
      '\nSE(beta_3) =', se_beta3,
      '\nSE(sigma) =', se_sigma)
```

```
SE(alpha) =  3.8635076855235617
SE(beta_1) = 0.18852774390326496
SE(beta_2) = 0.010751166266602833
SE(beta_3) = 0.012353905074660332
SE(sigma) = 3.8599340835674187
```

2.4. Constrained optimisation with `minimize`

`minimize` is also equipped with optimisation algorithms designed to deal with constraints. Let's say we believe that the intercept α must be between 5 and 6. The `trust-constr` (Trust-Region Constrained algorithm), `L-BFGS-B` (Limited memory BFGS with Box constraints), `TNC` (Truncated Newton, implemented in C), and `SLSQP` (Sequential Least SQuares Programming) methods can handle these constraints. See more information on these methods [here (https://docs.scipy.org/doc/scipy/reference/tutorial/optimize.html)](https://docs.scipy.org/doc/scipy/reference/tutorial/optimize.html). They are implemented by calling `method='chosenMethod'` as an argument of `minimize`.

To define our constraint $5 < \alpha < 6$, we actually need to define a system of inequalities over all parameters. For instance

$$\begin{bmatrix} 5 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \leq \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \sigma \end{bmatrix} \leq \begin{bmatrix} 6 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Where only the first inequality could be binding.

This is how we implement it:

In [8]:

```
#Defining the constraint with LinearConstraint
from scipy.optimize import LinearConstraint
cons = LinearConstraint([[1, 0, 0, 0, 0],
                         [0, 0, 0, 0, 0],
                         [0, 0, 0, 0, 0],
                         [0, 0, 0, 0, 0],
                         [0, 0, 0, 0, 0]],
                        [5, 1, 1, 1, 1],
                        [6, 1, 1, 1, 1])
```

In [9]:

```
results_cons = opt.minimize(negLogLNorm, β_0, args=(data), method='trust-constr'
, constraints=[cons])
results_cons
```

```
/Users/arnauddyevre/anaconda3/lib/python3.7/site-packages/scipy/opti
mize/_trustregion_constr/projections.py:182: UserWarning: Singular J
acobian matrix. Using SVD decomposition to perform the factorization
s.
  warn('Singular Jacobian matrix. Using SVD decomposition to ' +
```

Out[9]:

```
 barrier_parameter: 0.1
 barrier_tolerance: 0.1
          cg_niter: 1000
      cg_stop_cond: 1
            constr: [array([5.06085597, 0.        , 0.        , 0.
, 0.        ])]
       constr_nfev: [0]
       constr_nhev: [0]
       constr_njev: [0]
    constr_penalty: 387488835.02700907
  constr_violation: 1.0
    execution_time: 5.97819709777832
               fun: 60549.83434885077
              grad: array([  -47.99841898, -6973.0668413 ,  1248.900
25464, -7944.83056641,
       -9909.02994303])
               jac: [array([[1, 0, 0, 0, 0],
       [0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0]])]
   lagrangian_grad: array([-1.82012631e-01, -6.97306684e+03,  1.2489
0025e+03, -7.94483057e+03,
       -9.90902994e+03])
           message: 'The maximum number of function evaluations is e
xceeded.'
            method: 'tr_interior_point'
              nfev: 6006
              nhev: 0
               nit: 1001
             niter: 1001
              njev: 0
        optimality: 9909.029943030768
            status: 0
           success: False
         tr_radius: 22.745232042144288
                 v: [array([ 4.78164063e+01, -0.00000000e+00,  2.198
39790e-15, -0.00000000e+00,
       -0.00000000e+00])]
                 x: array([ 5.06085597,  3.72206003, -1.23437819,
0.3781635 , 11.44280202])
```

3. MLE with 'statsmodels'

It can be painful to code every single aspect of the Maximum Likelihood Estimation, and to fetch every interesting result manually. Fortunately, the `statsmodels` (https://www.statsmodels.org/stable/index.html) package has an extensive catalogue of tools for statistical inference, including MLE. This package also allows us to generate a much richer set of outputs very easily: confidence intervals, p-statistics, pseudo-R^2 are all generated by default.

STATA and R users will feel at home with this package as estimations results are presented in a similar way. However, using the package proficiently may sometimes require a good grasp of object-oriented programming. The time invested in getting familiar with object-oriented programming is definitely worth it; the payoffs will be huge down the road, when implementing more sophisticated routines will require to manipulate classes.

3.1. Hacking the `GenericLikelihoodModel` class from `statsmodels`

Let's first reproduce our results above with `statsmodels` , this will give us a feel of what output is generated. For this, we will need to define a new model class that inherits from `statsmodels` ' `GenericLikelihood Model` 's attributes. This class uses a log-likelihood (defined ex-ante), and will be estimated by `statsmodels` algorithms. When defining a custom model relying on maximum likelihood, we need to respect the standard architecture of the `GenericLikelihoodModel` canon (see here (https://www.statsmodels.org/dev/examples/notebooks/generated/generic_mle.html) for an example).

In [12]:

```
from scipy.stats import norm
import statsmodels.api as sm
from statsmodels.base.model import GenericLikelihoodModel

# We first define a log-likelihood, which will be fed to our custom class `MLE_f
or_OLS`
def logL(y, X, β, σ):
    y_hat = X @ β
    return norm(y_hat, σ).logpdf(y).sum()    # This is a shorter encoding of the
 log likelihood than the one above

# Now we build an MLE solver, using the 'GenericLikelihoodModel' class
class MLE_for_OLS(GenericLikelihoodModel):
    def __init__(self, endog, exog, **kwargs):
        super(MLE_for_OLS, self).__init__(endog, exog, **kwargs) # When we don't
know how many variable arguments can be passed on to the function, we use *args
                                                                  # We use **kwargs ins
tead when we want a named list of arguments (a dictionary)
    def nloglikeobs(self, params):
        σ = params[-1]
        β = params[:-1]
        ll = logL(self.endog, self.exog, β, σ)
        return -ll

    def fit(self, start_params = None, maxiter = 10000, maxfun = 10000, **kwargs
):
        # We need to add the σ to the list of parameters to be estimated
        self.exog_names.append('σ')
        if start_params == None:                              # Default starting v
alues, if none specified
            start_params = np.append(np.zeros(self.exog.shape[1]), 1)
        return super(MLE_for_OLS, self).fit(start_params=start_params,
                                maxiter=maxiter, maxfun=maxfun,
                                **kwargs)
```

And we now print the results.

In [13]:

```
sm_ols_manual = MLE_for_OLS( y, X).fit()
print(sm_ols_manual.summary())
```

```
Optimization terminated successfully.
         Current function value: 6.014743
         Iterations: 654
         Function evaluations: 1072
                             MLE_for_OLS Results
==============================================================================
Dep. Variable:                      y   Log-Likelihood:
-6014.7
Model:                    MLE_for_OLS   AIC:
1.204e+04
Method:            Maximum Likelihood   BIC:
1.206e+04
Date:                Mon, 23 Sep 2019
Time:                        12:04:58
No. Observations:                1000
Df Residuals:                     996
Df Model:                           3
==============================================================================
                 coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
const         -2.8253      6.241     -0.453      0.651     -15.058
9.408
x1             9.7120      0.260     37.282      0.000       9.201
10.223
x2            -0.5824      0.054    -10.760      0.000      -0.689
-0.476
x3             0.2544      0.013     19.408      0.000       0.229
0.280
σ             99.0605      2.227     44.476      0.000      94.695
103.426
==============================================================================
```

These results are very close to the ones we have found via manual encoding of MLE, and with OLS.

3.2. Exporting results to $\LaTeX$

The `as_latex()` command allows you to export your results in a neatly formatted TeX table. See the results of the last output when we write `print(sm_ols_manual.summary().as_latex())` . You will need the `booktabs` package for the table to compile in TeX.

In [14]:

```
print(sm_ols_manual.summary().as_latex())
```

```
\begin{center}
\begin{tabular}{lclc}
\toprule
\textbf{Dep. Variable:}    &         y          & \textbf{  Log-Like
lihood:    } &   -6014.7   \\
\textbf{Model:}            &    MLE\_for\_OLS    & \textbf{  AIC:
} & 1.204e+04   \\
\textbf{Method:}           & Maximum Likelihood & \textbf{  BIC:
} & 1.206e+04   \\
\textbf{Date:}             &  Mon, 23 Sep 2019  & \textbf{
} &             \\
\textbf{Time:}             &      12:05:33      & \textbf{
} &             \\
\textbf{No. Observations:} &         1000       & \textbf{
} &             \\
\textbf{Df Residuals:}     &          996       & \textbf{
} &             \\
\bottomrule
\end{tabular}
\begin{tabular}{lcccccc}
               & \textbf{coef} & \textbf{std err} & \textbf{z} & \te
xtbf{P$> |$z$|$} & \textbf{[0.025} & \textbf{0.975]}  \\
\midrule
\textbf{const} &      -2.8253  &        6.241     &    -0.453  &
0.651        &     -15.058    &        9.408     \\
\textbf{x1}    &       9.7120  &        0.260     &    37.282  &
0.000        &       9.201    &       10.223     \\
\textbf{x2}    &      -0.5824  &        0.054     &   -10.760  &
0.000        &      -0.689    &       -0.476     \\
\textbf{x3}    &       0.2544  &        0.013     &    19.408  &
0.000        &       0.229    &        0.280     \\
\textbf{σ}     &      99.0605  &        2.227     &    44.476  &
0.000        &      94.695    &      103.426     \\
\bottomrule
\end{tabular}
%\caption{MLE_for_OLS Results}
\end{center}
```

Dep. Variable:	y	**Log-Likelihood:**	-6014.7
Model:	MLE_for_OLS	**AIC:**	1.204e+04
Method:	Maximum Likelihood	**BIC:**	1.206e+04
Date:	Thu, 05 Sep 2019		
Time:	15:31:09		
No. Observations:	1000		
Df Residuals:	996		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-2.8255	6.242	-0.453	0.651	-15.060	9.409
x1	9.7119	0.261	37.282	0.000	9.201	10.223
x2	-0.5823	0.054	-10.757	0.000	-0.688	-0.476
x3	0.2545	0.013	19.410	0.000	0.229	0.280
σ	99.0619	2.227	44.475	0.000	94.696	103.427

We can also print several estimation results next to each other, as is standard in academic papers. To this end, we use the `summary_col` command.

In [15]:

```
from statsmodels.iolib.summary2 import summary_col

# We define an empty list, it will contain the regression outputs generated by o
ur MLE estimator
results = []

# We add one regressor at a time, and store the outputa in 'results'
for i in range(1, 5):
    col = MLE_for_OLS(y, X.T[0:i].T).fit()
    results.append(col)
```

```
Optimization terminated successfully.
         Current function value: 7.676231
         Iterations: 204
         Function evaluations: 398
Optimization terminated successfully.
         Current function value: 6.332248
         Iterations: 312
         Function evaluations: 554
Optimization terminated successfully.
         Current function value: 6.167039
         Iterations: 571
         Function evaluations: 968
Optimization terminated successfully.
         Current function value: 6.014743
         Iterations: 654
         Function evaluations: 1072
```

In [16]:

```
# We now use the functionalities of 'summary_col' to get a nicely formated table
summary = summary_col(results = results,
                      stars = True,
                      model_names = ['intercept', '2 variables', '3 variables',
'full'],
                      info_dict = {'Observations': lambda x: f"{int(x.nobs):d}"
},
                      float_format='%0.3f')
summary.add_title('OLS by MLE')
print(summary)
```

```
                        OLS by MLE
==========================================================
             intercept  2 variables 3 variables    full
----------------------------------------------------------
const        882.999*** 58.053***   31.344***   -2.825
             (16.501)   (8.258)     (7.129)     (6.241)
x1                      16.631***   11.724***   9.712***
                        (0.142)     (0.276)     (0.260)
x2                                  -1.082***   -0.582***
                                    (0.055)     (0.054)
x3                                              0.254***
                                                (0.013)
σ            521.804*** 136.089***  115.365***  99.061***
             (11.668)   (3.043)     (2.580)     (2.227)
Observations 1000       1000        1000        1000
==========================================================
Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01
```

And the TeX version, given by `print(summary.as_latex())`

Table 1: OLS by MLE

	intercept	2 variables	3 variables	full
const	882.999***	58.053***	31.344***	-2.825
	(16.501)	(8.258)	(7.129)	(6.241)
x1		16.631***	11.724***	9.712***
		(0.142)	(0.276)	(0.260)
x2			-1.082***	-0.582***
			(0.055)	(0.054)
x3				0.254***
				(0.013)
σ	521.804***	136.089***	115.365***	99.061***
	(11.668)	(3.043)	(2.580)	(2.227)
Observations	1000	1000	1000	1000

4. Discrete choice models with 'statsmodels'

The strength of `statmodels` lies in the breadth of its model library. All necessary tools for performing discrete choice analysis such as logit, probit, multinomial logit, Poisson or negative binomial come in canned commands.

We just show an application of the negative binomial model, as an example. All other models mentioned above are similarly implemented and we redirect you to the `statsmodels` documentation (https://www.statsmodels.org/dev/examples/notebooks/generated/discrete_choice_overview.html) on discrete choice models for more information.

Negative binomial regression is used to model count variables when the dependent variable is overdispersed. It is ideal when the dependent variable has a an excess count of zeros for instance. The probability mass function (PMF) of the negative binomial distribution gives us the probability that in a sequence of Bernoulli trials, k successes have occured when r failures have occured. In other words, if a Bernoulli trial with probability p of success has been repeated until r failures have occured, the number of successes X will be negative Binomial:

$$X \sim \mathrm{NB}(r, p)$$

and the PMF is

$$f(k|r,p) \equiv \Pr(X = k) = \binom{k+r-1}{k}(1-p)^r p^k$$

We first import a native 'statsmodels' dataset: "RAND". It was collected by the RAND corporation as part of a US country-wide health insurance study (1971-1986). The dependent variable is the number of visits to a doctor in a year, for a given individual. We will try to explain the number of visits by using variables such as insurance coverage, self-rated health, and number of chronic diseases.

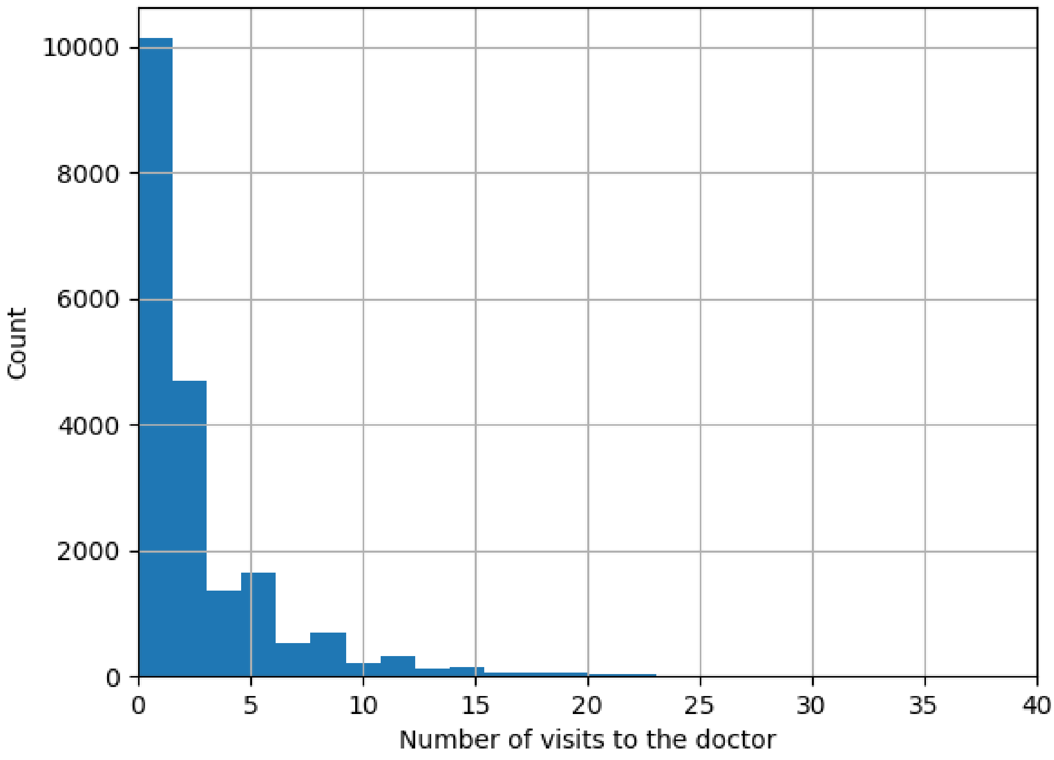
In [17]:

```
# Whole data
data = sm.datasets.randhie.load(as_pandas = False)

# Defining the set of exogenous variables, and adding an intercept
exog = data.exog.view(float).reshape(len(data.exog), -1)
exog = sm.add_constant(exog, prepend = False)
```

Our dependent variable is densely distributed around 0, which makes it a good candidate for the negative binomial model.

In [18]:

```
plt.figure(2)
plt.hist(data.endog, bins=50)
plt.xlim(xmin=0, xmax=40)
plt.grid()
plt.xlabel('Number of visits to the doctor')
plt.ylabel('Count')
plt.show()
```



We now fit the model:

In [19]:

```
results_NBin = sm.NegativeBinomial(data.endog, exog).fit()
results_NBin.summary()
```

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 2.148770
         Iterations: 35
         Function evaluations: 39
         Gradient evaluations: 39

/Users/arnauddyevre/anaconda3/lib/python3.7/site-packages/statsmodel
s/base/model.py:512: ConvergenceWarning: Maximum Likelihood optimiza
tion failed to converge. Check mle_retvals
  "Check mle_retvals", ConvergenceWarning)
```

Out[19]:

NegativeBinomial Regression Results

Dep. Variable:	y	**No. Observations:**	20190
Model:	NegativeBinomial	**Df Residuals:**	20180
Method:	MLE	**Df Model:**	9
Date:	Mon, 23 Sep 2019	**Pseudo R-squ.:**	0.01845
Time:	12:13:48	**Log-Likelihood:**	-43384.
converged:	False	**LL-Null:**	-44199.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
x1	-0.0579	0.006	-9.515	0.000	-0.070	-0.046
x2	-0.2678	0.023	-11.802	0.000	-0.312	-0.223
x3	0.0412	0.004	9.938	0.000	0.033	0.049
x4	-0.0381	0.003	-11.216	0.000	-0.045	-0.031
x5	0.2691	0.030	8.985	0.000	0.210	0.328
x6	0.0382	0.001	26.080	0.000	0.035	0.041
x7	-0.0441	0.020	-2.201	0.028	-0.083	-0.005
x8	0.0173	0.036	0.478	0.632	-0.054	0.088
x9	0.1782	0.074	2.399	0.016	0.033	0.324
const	0.6635	0.025	26.786	0.000	0.615	0.712
alpha	1.2930	0.019	69.477	0.000	1.256	1.329

Exercises

E.1. Poisson model with the RAND data

- Implement a Poisson model (`sm.Poisson( , )`) on the RAND data

Solution here (https://www.statsmodels.org/dev/examples/notebooks/generated/discrete_choice_overview.html?highlight=poisson)

- Compare the results to those generated by the negative binomial model and display them side by side

E.2. QuantEcon - MLE exercise 1 (https://lectures.quantecon.org/py/mle.html)

Solution at the bottom of the QuantEcon page