# Which prisoners are likely to reoffend?

Amal FLIMINE, Alex FRANÇOIS

ENAC

May 7, 2020

# Table of contents

# Questions of interest

- Can we predict who will be more likely to reoffend?
- Among the repeat offenders, what was the duration before they returned to jail?

# Dataset

- A Wooldridge dataset on North Carolina prisoners
- Cross sectional data
- Total observations : 1445

# Dependant variables

- **CENS** : 1 if duration right censored, 0 if not (*i.e* if repeat offense before the end of follow period)

  - *Binary variable*

- **LDURAT** = log(durat) where durat = *max*(time until return, follow) and is a,

  - *Continuous variable (months)*

# Descriptive variables

| Variables | Type | Unit | Expected |
|---|---|---|---|
| BLACK | Binary | - | - |
| ALCOHOL | Binary | - | - |
| DRUGS | Binary | - | - - |
| SUPER | Binary | - | + |
| MARRIED | Binary | - | + |
| FELON | Binary | - | - |
| WORKPRG | Binary | - | + |
| PROPERTY | Binary | - | - - |
| PERSON | Binary | - | - |
| PRIORS | Count | - | - |
| EDUC | Count | years | ++ |
| RULES | Count | - | - |
| AGE | Count | months | ? |
| TSERVED | Continuous | months | ? |

Table: Explanatory variables for LDURAT (or CENS)

# Binary variables

| **Variables** | **Percentage** |
|---|---|
| BLACK = 1 | 49% |
| ALCOHOL = 1 | 21% |
| DRUGS = 1 | 24% |
| SUPER = 1 | 69% |
| MARRIED = 1 | 25.5% |
| FELON =1 | 31% |
| WORKPRG =1 | 46% |
| PROPERTY =1 | 25% |
| PERSON =1 | 5% |

Table: Distribution of dummies.

# Discrete variables

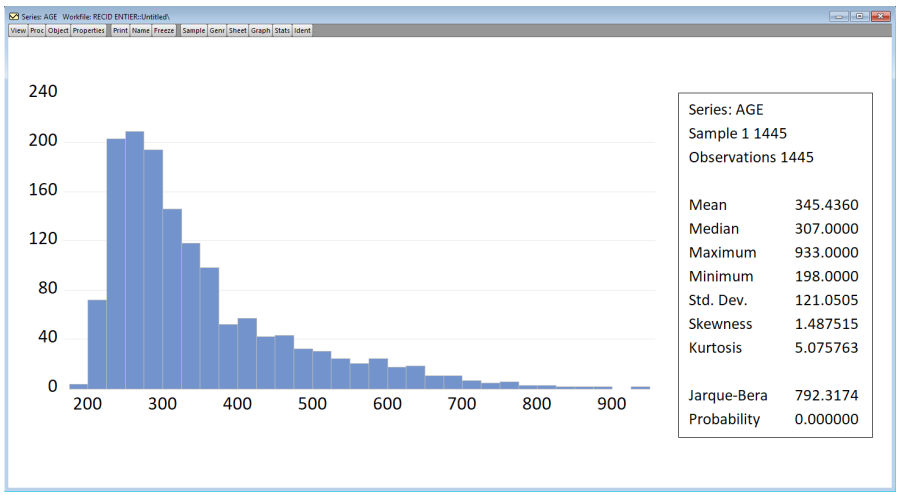


Figure: AGE histogram

KT>3, SK>0

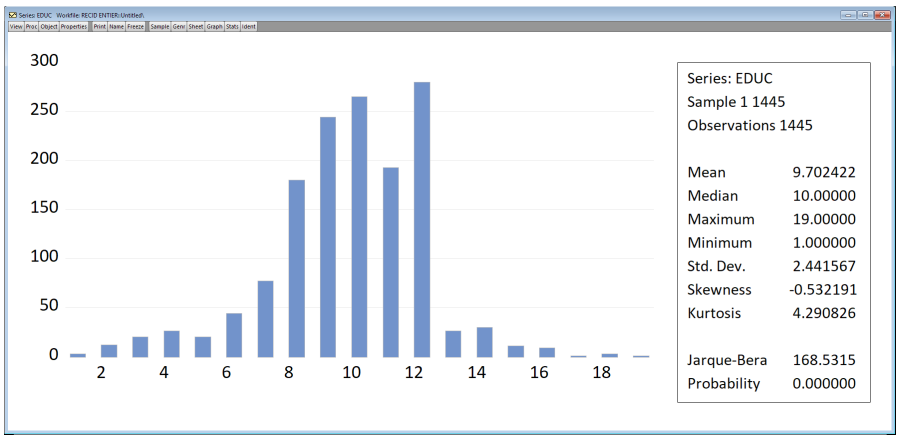


Figure: EDUC histogram

Few values above 16 years

# Discrete variables

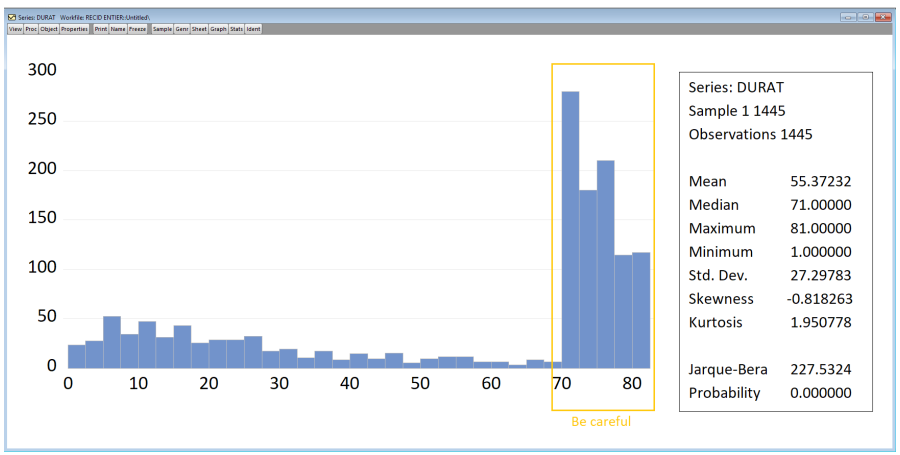


Figure: DURAT histogram

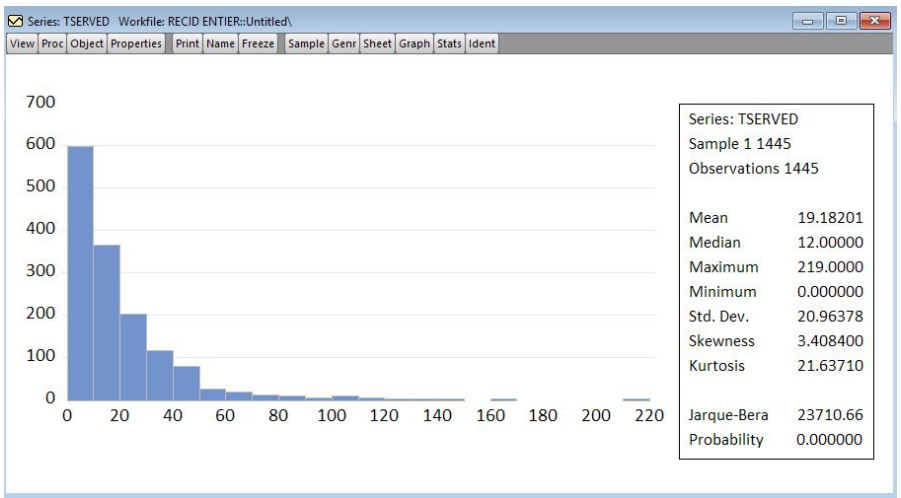


Figure: TSERVED histogram

Be careful (links between DURAT and CENS)

# Discrete variables

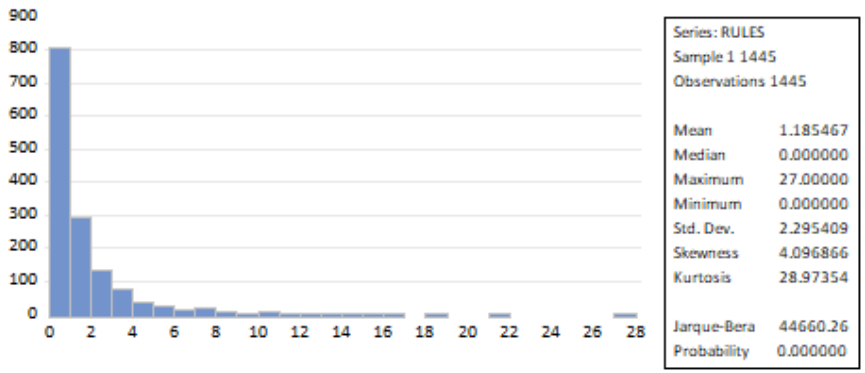


Figure: RULES histogram

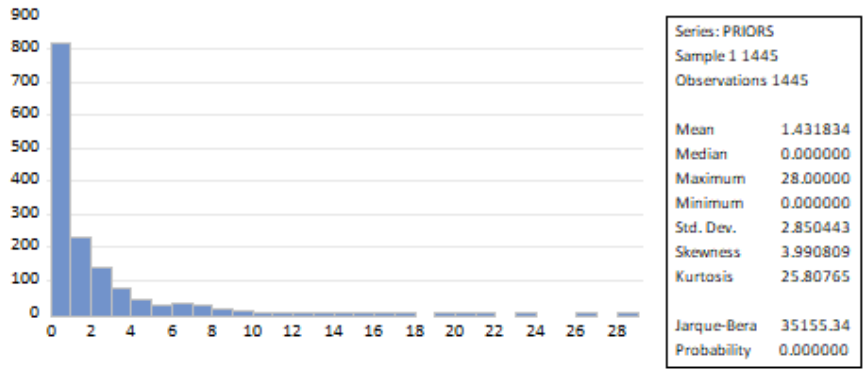


Figure: PRIORS histogram

Correlation between explanatory variables?

# Analysis of correlation matrix

- + : RULES/TSERVED (r=0.51) AGE/PRIORS (r=0.42) TSERVED/PRIORS (r=0.17)
- - : AGE/EDUC (r=-0.27) AGE/RULES (r=-0.15)

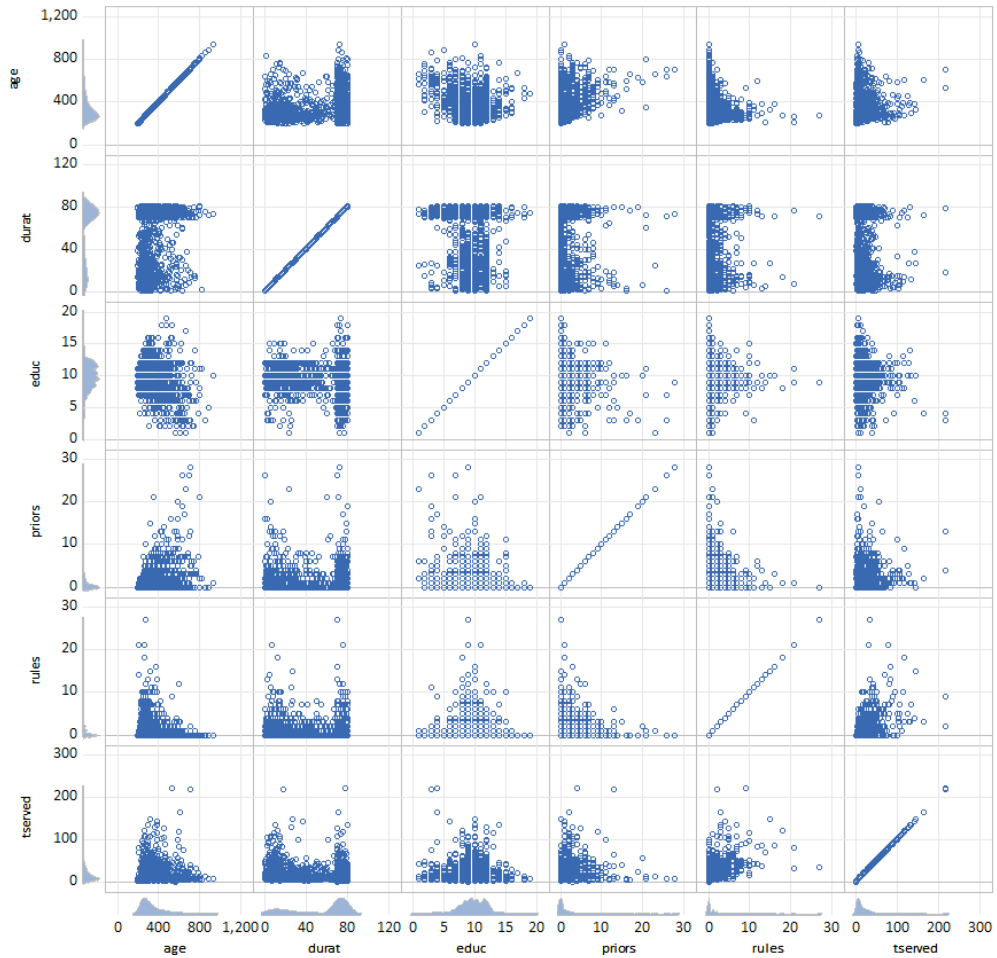


Figure: Correlation graph for non binary variables

# Choice of variables

- Some unbalanced data : insignificant explanatory variables.
- We won't use FOLLOW neither regress (LDURAT on CENS).
- For duration analysis, we should consider uncensored data (552 obs for CENS=0).

# Equation

- We used Probit regression to model CENS

$$P(\text{CENS}=1)=\beta_0+\beta_1*AGE+\beta_2*ALCOHOL+\beta_3*BLACK+\beta_4*DRUGS+\beta_5*EDUC+\beta_6*FELON+\beta_7*MARRIED+\beta_8*PERSON+\beta_9*PRIORS+\beta_{10}*PROPERTY+\beta_{11}*RULES+\beta_{12}*SUPER+\beta_{13}*TSERVED+\beta_{14}*WORKPRG$$

# Linear regression (Probit model)

Dependent Variable: CENS
Method: ML - Binary Probit (Newton-Raphson / Marquardt steps)
Date: 05/06/20 Time: 08:38
Sample: 1 1445
Included observations: 1445
Convergence achieved after 3 iterations
Coefficient covariance computed using observed Hessian

| Variable | Coefficient | Std. Error | z-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.093467 | 0.219336 | -0.426138 | 0.6700 |
| AGE | 0.002180 | 0.000372 | 5.855325 | 0.0000 |
| ALCOHOL | -0.347595 | 0.089724 | -3.874058 | 0.0001 |
| BLACK | -0.357747 | 0.071872 | -4.977528 | 0.0000 |
| DRUGS | -0.205929 | 0.082574 | -2.493858 | 0.0126 |
| EDUC | 0.021043 | 0.015457 | 1.361409 | 0.1734 |
| FELON | 0.484679 | 0.140777 | 3.442877 | 0.0006 |
| MARRIED | 0.134081 | 0.085110 | 1.575383 | 0.1152 |
| PERSON | -0.126576 | 0.204430 | -0.619163 | 0.5358 |
| PRIORS | -0.070520 | 0.014025 | -5.028109 | 0.0000 |
| PROPERTY | -0.322962 | 0.137889 | -2.342195 | 0.0192 |
| RULES | -0.030568 | 0.017925 | -1.705307 | 0.0881 |
| SUPER | 0.021196 | 0.079090 | 0.268001 | 0.7887 |
| TSERVED | -0.009255 | 0.002261 | -4.093078 | 0.0000 |
| WORKPRG | -0.085271 | 0.074445 | -1.145425 | 0.2520 |

| | | | |
|---|---|---|---|
| McFadden R-squared | 0.079516 | Mean dependent var | 0.617993 |
| S.D. dependent var | 0.486046 | S.E. of regression | 0.462404 |
| Akaike info criterion | 1.245075 | Sum squared resid | 305.7585 |
| Schwarz criterion | 1.299841 | Log likelihood | -884.5664 |
| Hannan-Quinn criter. | 1.265515 | Deviance | 1769.133 |
| Restr. deviance | 1921.960 | Restr. log likelihood | -960.9800 |
| LR statistic | 152.8272 | Avg. log likelihood | -0.612157 |
| Prob(LR statistic) | 0.000000 | | |

| | | | |
|---|---|---|---|
| Obs with Dep=0 | 552 | Total obs | 1445 |
| Obs with Dep=1 | 893 | | |

Figure: CENS regression with Probit

# Classical assumptions (OLS regression)

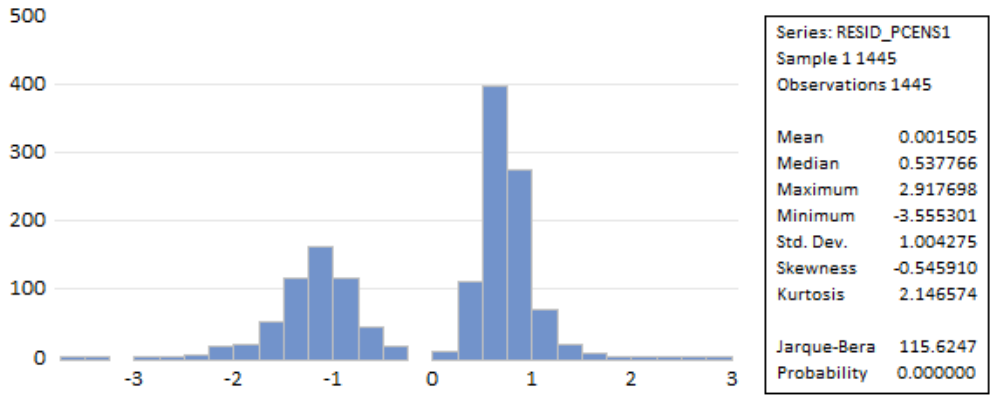


Figure: Residuals histogram

Does normality really matter here?

# Classical assumptions (OLS regression)

Heteroskedasticity Test: Breusch-Pagan-Godfrey
Null hypothesis: Homoskedasticity

| | | | |
|---|---|---|---|
| F-statistic | 5.251689 | Prob. F(14,1430) | 0.000 |
| Obs*R-squared | 70.66178 | Prob. Chi-Square(14) | 0.000 |
| Scaled explained SS | 20.03192 | Prob. Chi-Square(14) | 0.129 |

Figure: BPG's test for heteroskedasticity.

Evidence of heteroskedasticity at 99%.

# Classical assumptions (OLS regression)

Breusch-Godfrey Serial Correlation LM Test:
Null hypothesis: No serial correlation at up to 2 lags

| | | | |
|---|---|---|---|
| F-statistic | 0.124019 | Prob. F(2,1428) | 0.8834 |
| Obs*R-squared | 0.250947 | Prob. Chi-Square(2) | 0.8821 |

Figure: BG's LM test for serial correlation

Can't reject the null "no serial correlation"

# Forecast

Expectation-Prediction Evaluation for Binary Specification
Equation: UNTITLED
Date: 05/07/20 Time: 10:55
Success cutoff: C = 0.5

| | Estimated Equation | | | Constant Probability | | |
|---|---|---|---|---|---|---|
| | Dep=0 | Dep=1 | Total | Dep=0 | Dep=1 | Total |
| P(Dep=1)<=C | 179 | 104 | 283 | 0 | 0 | 0 |
| P(Dep=1)>C | 373 | 789 | 1162 | 552 | 893 | 1445 |
| Total | 552 | 893 | 1445 | 552 | 893 | 1445 |
| Correct | 179 | 789 | 968 | 0 | 893 | 893 |
| % Correct | 32.43 | 88.35 | 66.99 | 0.00 | 100.00 | 61.80 |
| % Incorrect | 67.57 | 11.65 | 33.01 | 100.00 | 0.00 | 38.20 |
| Total Gain* | 32.43 | -11.65 | 5.19 | | | |
| Percent Gain** | 32.43 | NA | 13.59 | | | |

Figure: Expectation-Prediction Evaluation of CENS

# New set of data

- Uncensored observations : n=552
- Dependant variable : LDURAT corresponds to a duration (1-77) with mean=23.9mo (low).

| **Variables** | **Full Data** | **Uncensored data** | |
|---|---|---|---|
| BLACK = 1 | 49% | 56% | ↑ |
| ALCOHOL = 1 | 21% | 24% | ↑ |
| DRUGS = 1 | 24% | 27% | ↑ |
| MARRIED = 1 | 25.5% | 21% | ↓ |
| WORKPRG =1 | 46% | 48% | ↑ |
| PROPERTY =1 | 25% | 29% | ↑ |

Table: Some changes in statistics.

# Equation

- We used OLS regression to model LDURAT :

$$\text{LDURAT} = \beta_0 + \beta_1 * AGE + \beta_2 * AGE^2 + \beta_3 * TSERVED + \beta_4 * TSERVED^2 + \beta_5 * PRIORS + \beta_6 * RULES + \beta_7 * ALCOHOL + \beta_8 * BLACK + \beta_9 * DRUGS + \beta_{10} * (EDUC < 9) + \beta_{11} * MARRIED + \beta_{12} * PROPERTY + \beta_{13} * WORKPRG + \beta_{14} * SUPER + \beta_{15} * PERSON + \beta_{16} * FELON$$

# Linear regression (OLS)

Equation: EQLDURAT Workfile: RECID (CENS=0)::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LDURAT
Method: Least Squares
Date: 05/06/20 Time: 15:04
Sample: 1 552
Included observations: 552

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 1.980313 | 0.401575 | 4.931368 | 0.0000 |
| AGE | 0.005870 | 0.002152 | 2.727849 | 0.0066 |
| AGE^2 | -6.70E-06 | 2.45E-06 | -2.730151 | 0.0065 |
| TSERVED | -0.019604 | 0.004627 | -4.236440 | 0.0000 |
| TSERVED^2 | 0.000100 | 3.00E-05 | 3.345809 | 0.0009 |
| PRIORS | -0.047556 | 0.013908 | -3.419291 | 0.0007 |
| RULES | -0.006362 | 0.018874 | -0.337081 | 0.7362 |
| ALCOHOL | -0.263157 | 0.096045 | -2.739936 | 0.0064 |
| BLACK | -0.005303 | 0.081203 | -0.065304 | 0.9480 |
| DRUGS | 0.004565 | 0.087659 | 0.052073 | 0.9585 |
| EDUC<9 | 0.266005 | 0.090333 | 2.944697 | 0.0034 |
| MARRIED | 0.185776 | 0.098784 | 1.880627 | 0.0606 |
| PROPERTY | -0.178164 | 0.147684 | -1.206387 | 0.2282 |
| WORKPRG | 0.047314 | 0.085893 | 0.550840 | 0.5820 |
| SUPER | 0.044691 | 0.086831 | 0.514684 | 0.6070 |
| PERSON | 0.000479 | 0.220858 | 0.002169 | 0.9983 |
| FELON | 0.363036 | 0.153219 | 2.369396 | 0.0182 |

| | | | |
|---|---|---|---|
| R-squared | 0.121272 | Mean dependent var | 2.824414 |
| Adjusted R-squared | 0.094992 | S.D. dependent var | 0.930001 |
| S.E. of regression | 0.884728 | Akaike info criterion | 2.623239 |
| Sum squared resid | 418.7674 | Schwarz criterion | 2.756084 |
| Log likelihood | -707.0139 | Hannan-Quinn criter. | 2.675144 |
| F-statistic | 4.614667 | Durbin-Watson stat | 2.005672 |
| Prob(F-statistic) | 0.000000 | | |

Figure: LDURAT regression with OLS

# Linear regression (OLS) - Wald Test

Wald Test:
Equation: EQLDURAT

| Test Statistic | Value | df | Probability |
|---|---|---|---|
| F-statistic | 0.146462 | (6, 535) | 0.9897 |
| Chi-square | 0.878769 | 6 | 0.9898 |

Figure: Wald Test

C(RULES)=C(BLACK)=C(DRUGS)=
C(WORKPRG)=C(SUPER)=C(PERSON)=0

Dependent Variable: LDURAT
Method: Least Squares
Date: 05/06/20 Time: 14:57
Sample: 1 552
Included observations: 552

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 2.176375 | 0.403082 | 5.399340 | 0.0000 |
| AGE | 0.006512 | 0.002060 | 3.161143 | 0.0017 |
| AGE^2 | -7.21E-06 | 2.38E-06 | -3.024038 | 0.0026 |
| TSERVED | -0.019924 | 0.004125 | -4.830342 | 0.0000 |
| TSERVED^2 | 0.000101 | 2.91E-05 | 3.483438 | 0.0005 |
| PRIORS | -0.047537 | 0.013889 | -3.422588 | 0.0007 |
| ALCOHOL | -0.275162 | 0.094023 | -2.926539 | 0.0036 |
| EDUC | -0.023687 | 0.018678 | -1.268162 | 0.2053 |
| MARRIED | 0.190588 | 0.098200 | 1.940808 | 0.0528 |
| PROPERTY | -0.188358 | 0.129062 | -1.459441 | 0.1450 |
| FELON | 0.372246 | 0.140019 | 2.658536 | 0.0081 |

| | | | |
|---|---|---|---|
| R-squared | 0.107073 | Mean dependent var | 2.824414 |
| Adjusted R-squared | 0.090567 | S.D. dependent var | 0.930001 |
| S.E. of regression | 0.886888 | Akaike info criterion | 2.617530 |
| Sum squared resid | 425.5344 | Schwarz criterion | 2.703488 |
| Log likelihood | -711.4382 | Hannan-Quinn criter. | 2.651116 |
| F-statistic | 6.487229 | Durbin-Watson stat | 2.016991 |
| Prob(F-statistic) | 0.000000 | | |

Figure: Regression obtained

# Classical assumptions tests - Normality of $\hat{u}$

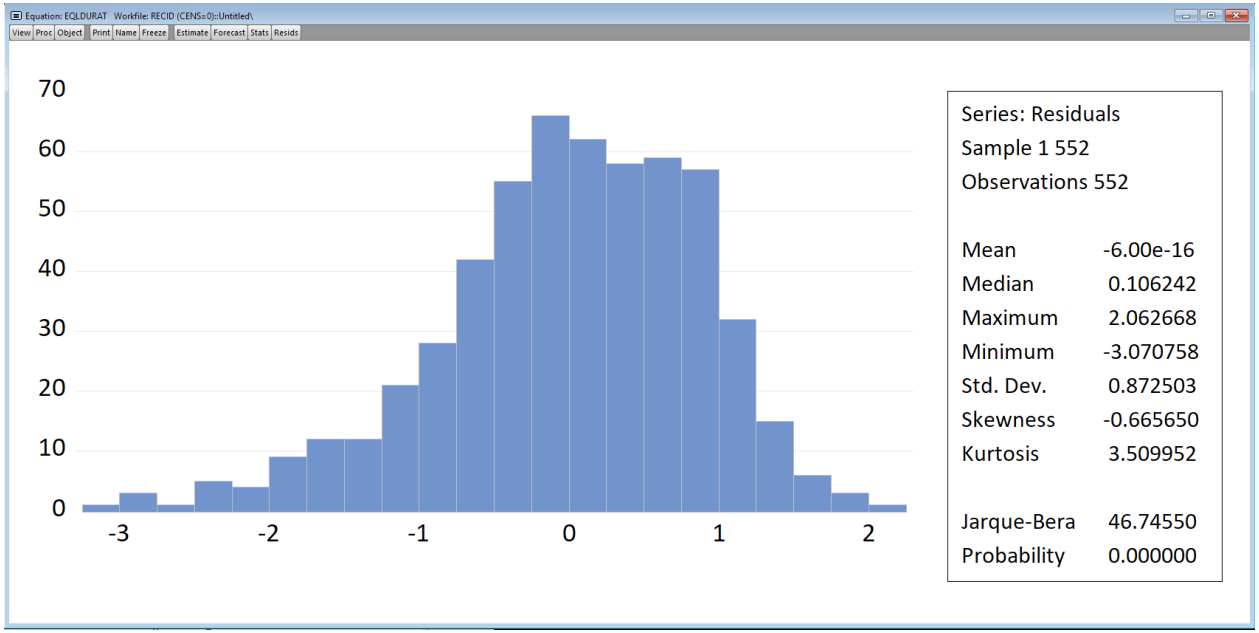


Figure: Residuals histogram

Reject the null of normality at 99% level

# Classical assumptions tests - Heteroskedasticity and serial correlation

Heteroskedasticity Test: Breusch-Pagan-Godfrey
Null hypothesis: Homoskedasticity

| | | | |
|---|---|---|---|
| F-statistic | 2.175581 | Prob. F(10,541) | 0.0179 |
| Obs*R-squared | 21.34000 | Prob. Chi-Square(10) | 0.0188 |
| Scaled explained SS | 25.72445 | Prob. Chi-Square(10) | 0.0041 |

Figure: BPG Test

Reject the null of homoskedasticity at 98% level

Breusch-Godfrey Serial Correlation LM Test:
Null hypothesis: No serial correlation at up to 2 lags

| | | | |
|---|---|---|---|
| F-statistic | 3.028455 | Prob. F(2,539) | 0.0492 |
| Obs*R-squared | 6.134064 | Prob. Chi-Square(2) | 0.0466 |

Figure: BP autocorrelation LM Test

Reject the null of autocorrelation at 95% level

# Improving the model

Equation: EQLDURAT Workfile: RECID (CENS=0)::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LDURAT
Method: Least Squares
Date: 05/06/20 Time: 15:22
Sample: 1 552
Included observations: 552
HAC standard errors & covariance (Bartlett kernel, Newey-West fixed bandwidth = 6.0000)

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 1.921354 | 0.321810 | 5.970458 | 0.0000 |
| AGE | 0.006322 | 0.001713 | 3.689843 | 0.0002 |
| AGE^2 | -7.14E-06 | 1.99E-06 | -3.579296 | 0.0004 |
| TSERVED | -0.019656 | 0.003499 | -5.617796 | 0.0000 |
| TSERVED^2 | 9.99E-05 | 2.05E-05 | 4.867146 | 0.0000 |
| PRIORS | -0.048572 | 0.017619 | -2.756803 | 0.0060 |
| ALCOHOL | -0.263671 | 0.093105 | -2.831986 | 0.0048 |
| EDUC<9 | 0.272113 | 0.084791 | 3.209233 | 0.0014 |
| MARRIED | 0.194984 | 0.091320 | 2.135189 | 0.0332 |
| PROPERTY | -0.179079 | 0.112611 | -1.590245 | 0.1124 |
| FELON | 0.376988 | 0.126428 | 2.981841 | 0.0030 |

| | | | |
|---|---|---|---|
| R-squared | 0.119829 | Mean dependent var | 2.824414 |
| Adjusted R-squared | 0.103559 | S.D. dependent var | 0.930001 |
| S.E. of regression | 0.880530 | Akaike info criterion | 2.603141 |
| Sum squared resid | 419.4553 | Schwarz criterion | 2.689099 |
| Log likelihood | -707.4669 | Hannan-Quinn criter. | 2.636727 |
| F-statistic | 7.365313 | Durbin-Watson stat | 2.013558 |
| Prob(F-statistic) | 0.000000 | Wald F-statistic | 10.78965 |
| Prob(Wald F-statistic) | 0.000000 | | |

Figure: Newey-West Standard Errors regression of LDURAT

Robust to auto-correlation and heteroskedasticity

# Limits of the model?

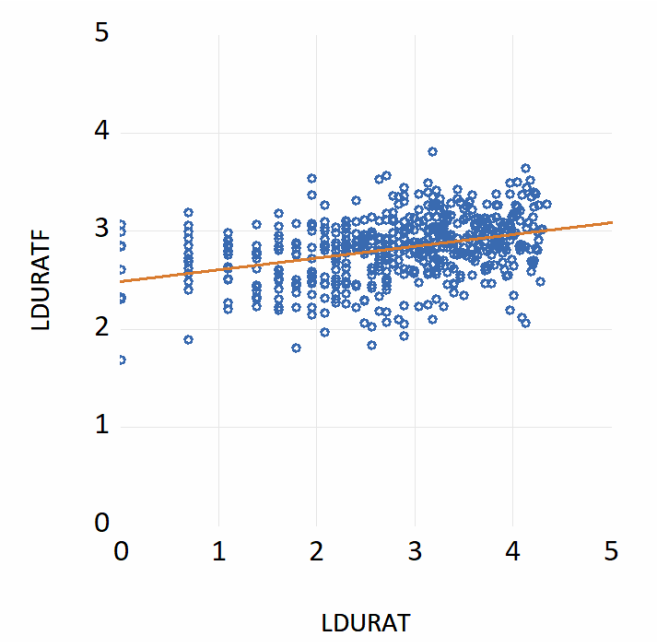


Figure: LDURATF against LDURAT

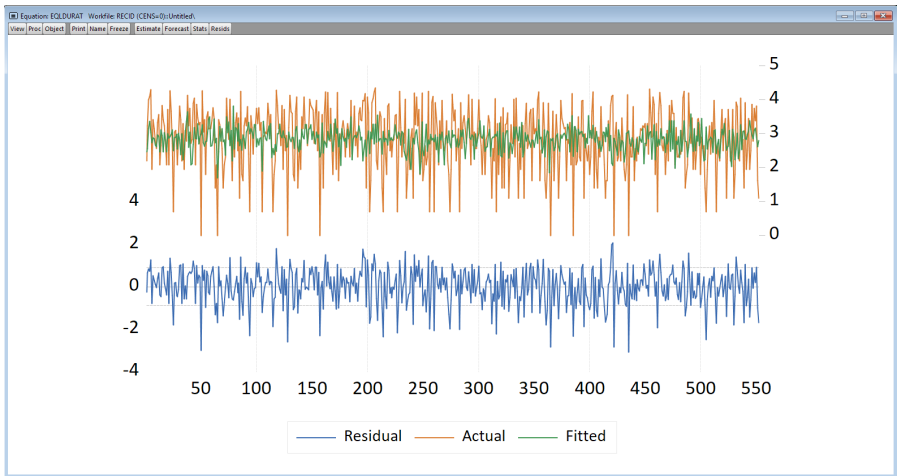


Figure: Actual, fitted, residual table

Reject the null of auto-correlation at 95% level

Struggle to have a nicely fitted regression.

# Marginal effects: actual v.s expected

| Variables | Expected | Model (CENS) | Model (LDURAT) + m.e |
|---|---|---|---|
| BLACK | - | - - | (insig) |
| ALCOHOL | - | - - | - - (30%) |
| DRUGS | - - | - - | (insig) |
| SUPER | + | (insig) | (insig) |
| MARRIED | + | + | + (19%) |
| FELON | - | ++ | ++ (46%) |
| WORKPRG | + | (insig) | (insig) |
| PROPERTY | - - | - | - (16%) |
| PERSON | - | (insig) | (insig) |
| PRIORS | - | - | - (5%) |
| EDUC | ++ | + | (insig) |
| RULES | - | - | - (2%/rule) |
| AGE | ? | + | + ( 7%/year) |
| TSERVED | ? | - | - - ( 24%/year) |

Table: Explanatory variables for LDURAT and CENS

Marginal effects for dummy variables: 100*($e^{\beta}$-1)

# Strengths and weaknesses of the models

## Model 1 (CENS) - Probit

- (-) Non normality of residuals?
- (-) Need other techniques to test for CAs
- (-) Only 32% of repeat offenders detected - Forecast isn't accurate !

- (+) Solid to model binary dependant variables.
- (+) Probabilities can be computed in various ways.

## Model 2 (LDURAT) - GLS

- (-) Non normality of the errors
- (-) Doesn't take into account censored data
- (-) Regression doesn't very fit well the data ($R^2 \leq 15\%$)

- (+) Heteroskedasticity and autocorrelation can be overcome with *i.e* Newey–West estimator
- (+) Simple model with qualitative marginal effects

# So, can we tell which inmates are more likely to reoffend?

- Difficulties in predicting both the reoffenders and non-reoffenders
- Results can differ from a model to another (*i.e* significance of BLACK)
- But some good influences found :

*Average profile of a likely repeat offender:*

*A black and aged person who didn't have much education, and previously had problems with drugs and alcohol. He didn't commit serious crimes (rather property infractions), but have many priors and transgressed many rules while incarcerated.*

# To overstep these problems we need

- More data (normality of u)
- More variables (criminality depends on a huge amount of parameters)
- Other kinds of regressions (censored regression model for LDURAT)

# Some references



Binary dependant variable models, 2019.


Monnery, B.
Prison, reentry and recidivism : micro-econometric applications.


Wooldridge, J. M.
*Introductory Econometrics: A Modern Approach*, 5th ed.
South-Western College Publishing, 2000.