How sporty are you ?

Guillaume Biehlmann - Florian Mariaud

ENAC, IENAC14 OPS-PREV

April 14, 2016

Table of contents

Our first main interrogations

- How much time people spend playing sports ?
- What brought people to begin sports ?
- What kind of sports are mostly practiced ?
- What makes one sportier than another ?

Table of contents

Our aims for this project :

- Collect data on people's sporty behaviour

Our aims for this project :

- Collect data on people's sporty behaviour
- Create an Econometrics model of this behaviour

Our aims for this project :

- Collect data on people's sporty behaviour
- Create an Econometrics model of this behaviour
- Interpret our data with EViews and try to modelize the results to find an answer to our problem

Table of contents

The survey

Database

- Use of a Google Form (in French)
- n = 400 answers collected in 5 days
- Majority of students (70 %)
- A total of k = 33 variables

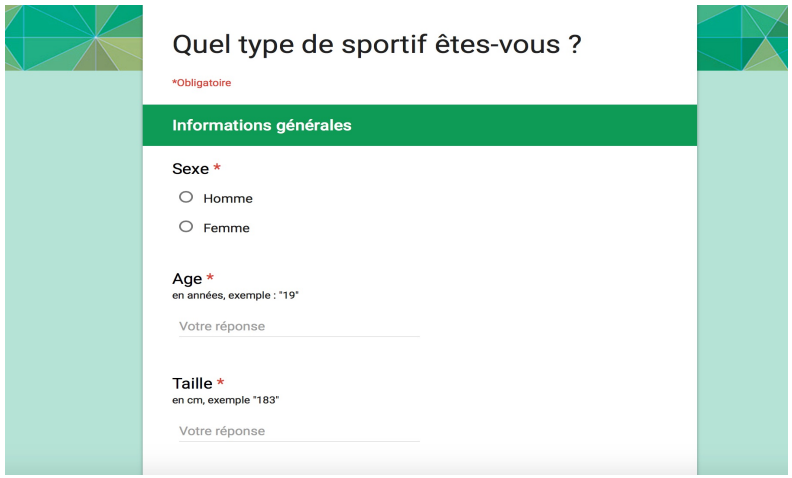

Figure 1: Our survey.

Table of contents

Age

- Majority of young adults (median = 21)
- Usually more concerned by sports
- Model might be biaised

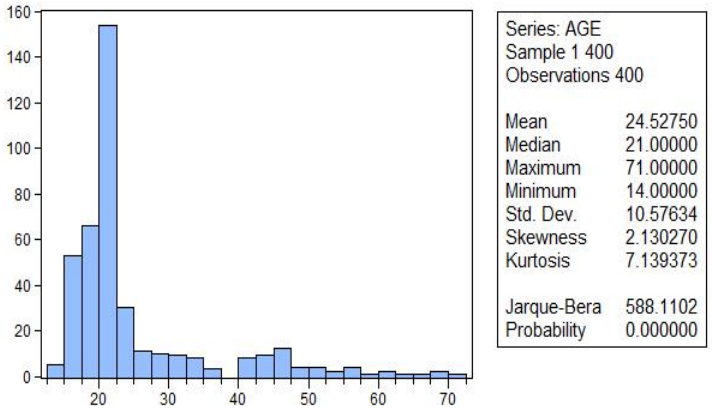

Figure 2: Age histogram.

Intro/Extrovert

- People had to rank themselves on a 10 point scale (10 = perfectly extrovert)
- JB test gives us a normal law at a 90 % level of confidence

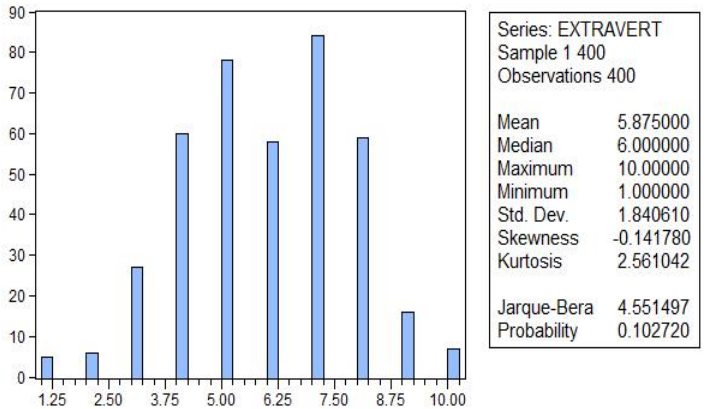

Figure 3: Introvert histogram

Cigarettes

- Majority of people do not smoke
- This variable will not be significant (biaised model)

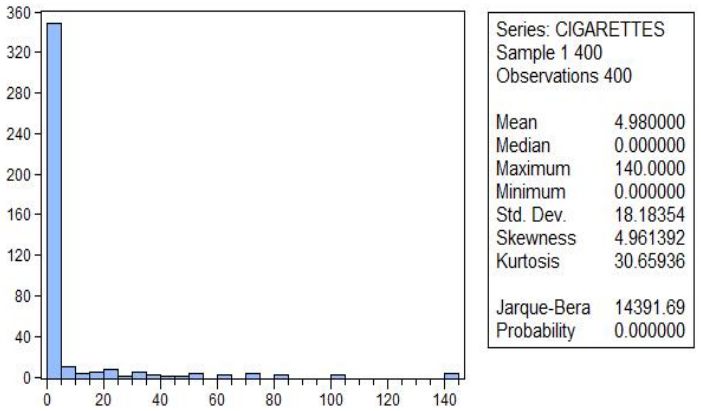

Figure 4: Cigarettes histogram

Table of contents

How would you describe a sporty person ?

- How often they practice sports per week ? $\longrightarrow$ **Frequency**
- How much time do they spend exercising ? $\longrightarrow$ **Sp_duration**
- How people judge the intensity of their sport ? $\longrightarrow$ **Intensity**
- How motivated they are ? $\longrightarrow$ **Motivation**
- Depending on their sport, how many calories do they burn per hour ? $\longrightarrow$ **Calories_hour**

Table of contents

Variables for the sport frequency model

How do we choose the variables :

- Analysis of the correlation matrix
- Suppression of some variables to avoid multicollinearity
- Non-significant variables (with p $>$ 0.15)
- Try to increase $\overline{R^2}$

Analysis of the correlation matrix

Variable	Correlated with	Coefficient
Frequency	Intensity	0.66
-	Motivation	0.59
-	Place_sp	0.64
-	Sp_duration	0.64
Age	Kids	0.82
-	Relationship	0.50
-	Salary	0.64
Kids	Relationship	0.53
-	Salary	0.54
Continue_competition	Competition	0.55
Height	Man	0.68
-	Weight	0.68

Table 1: Main significant coefficients in the correlation matrix

What will be our Y ?

Dependant variable

- We did not consider Calories_hour because the model was of lesser quality
- As Frequency, Intensity, Motivation, Sp_duration were correlated, we chose the one giving us the best explanatory model $\longrightarrow$ **Frequency will be our Y**

Table of contents

The first linear regression

Dependent Variable: FREQUENCY
Method: Least Squares
Date: 04/05/16 Time: 17:12
Sample (adjusted): 1 397
Included observations: 323 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
C	3.353057	1.078703	3.108417	0.0021
_SEG_SPORT	0.106705	0.037420	2.851531	0.0047
BEGIN_FAMILY	-0.015539	0.131198	-0.118438	0.9058
BEGIN_FRIENDS	-0.005985	0.140264	-0.042668	0.9660
BEGIN_IDK	-0.169769	0.237827	-0.713831	0.4760
BEGIN_PEOPLE	-0.339840	0.185767	-1.829389	0.0685
BEGIN_PRESSUREOFF	-0.123878	0.131739	-0.940324	0.3479
BEGIN_SCHOOL	0.025194	0.153876	0.163730	0.8701
BEGIN_TV	0.065926	0.194742	0.338530	0.7352
BEGIN_WORKOUT	0.010601	0.180904	0.058603	0.9533
BMI	-0.013914	0.021709	-0.640920	0.5221
BUDGET	0.007575	0.003085	2.455001	0.0147
CIGARETTES	0.001816	0.003291	0.551836	0.5815
COMPETITION	0.866104	0.170756	5.072162	0.0000
CONTINUE_BODYBUILD	0.377707	0.148935	2.536052	0.0118
CONTINUE_FAMILY	0.234551	0.154920	1.514014	0.1312
CONTINUE_FUN	0.185342	0.152978	1.211561	0.2268
CONTINUE_MEETPEOPLE	0.166625	0.150204	1.109323	0.2683
CONTINUE_PROGRESS	0.460159	0.144446	3.185692	0.0016
CONTINUE_TALENT	-0.005616	0.166270	-0.033775	0.9731
CONTINUE_TEAM_CLUB	-0.159519	0.157272	-1.014285	0.3114
CONTINUE_WORKOUT	-0.202755	0.140971	-1.438272	0.1516
DISTANCE	0.000102	0.008812	0.011605	0.9907
EXTRAVERT	0.062856	0.034003	1.848539	0.0657
IF_FREETIME	0.045709	0.137579	0.332237	0.7400
KIDS	0.005516	0.114939	0.047992	0.9618
LEISURE_TIME	-0.001132	0.005789	-0.195633	0.8451
MAN	0.186947	0.140974	1.326103	0.1860
OCCUPATION=1	-0.958459	0.796351	-1.203563	0.2299
OCCUPATION=2	-0.759426	0.882089	-0.860940	0.3901
OCCUPATION=3	-1.085153	0.784317	-1.383565	0.1677
OCCUPATION=4	-1.012163	0.793934	-1.274871	0.2035
OCCUPATION=5	-0.220820	1.033619	-0.213638	0.8310
OCCUPATION=6	0.728927	1.077643	0.676408	0.4994
OCCUPATION=7	-0.736204	1.025451	-0.717932	0.4734
OCCUPATION=8	0.720997	0.917538	0.785796	0.4327
OCCUPATION=9	-1.005781	0.818921	-1.228179	0.2205
RAISED=1	-0.189912	0.141490	-1.342233	0.1807
RELATIONSHIP	-0.173497	0.112127	-1.547323	0.1230
SALARY	3.05E-05	0.000108	0.282836	0.7775
SPORT_DIET	0.634698	0.148739	4.267203	0.0000
SPORT_TV	0.014427	0.023240	0.620797	0.5353
TRANSPORT=2	0.107145	0.278602	0.384582	0.7009
TRANSPORT=3	0.001792	0.173049	0.010356	0.9917
TRANSPORT=4	0.149853	0.326985	0.458288	0.6471
TRANSPORT=5	0.202821	0.255611	0.793473	0.4282
WEIGHT_WATCH	-0.281419	0.136426	-2.062790	0.0401
WHERE_LIVE=1	-0.004936	0.191854	-0.025728	0.9795
WHERE_LIVE=2	0.048767	0.138621	0.351803	0.7253
WORK	-0.011899	0.006428	-1.851194	0.0653
CALORIES_HOUR	-9.49E-05	0.000760	-0.124830	0.9008
MAIN_SP=0	-1.698353	0.480695	-3.533120	0.0005
MAIN_SP=1	-0.251906	0.371585	-0.677922	0.4984
MAIN_SP=2	-0.589704	0.335690	-1.756692	0.0801
MAIN_SP=3	-0.164294	0.407205	-0.403467	0.6869
MAIN_SP=4	0.471138	0.406679	1.158499	0.2477
MAIN_SP=5	-0.644317	0.490214	-1.314357	0.1899
MAIN_SP=6	-0.719250	0.503596	-1.428229	0.1544
MAIN_SP=7	-0.777336	0.695787	-1.117204	0.2649
MAIN_SP=8	0.644081	0.369653	1.742393	0.0826
MAIN_SP=9	-1.490861	0.817473	-1.823743	0.0693
MAIN_SP=10	-1.669878	0.752693	-2.218538	0.0274
MAIN_SP=11	0.122575	0.407847	0.300542	0.7640

R-squared	0.636819	Mean dependent var	3.049536
Adjusted R-squared	0.550214	S.D. dependent var	1.454493
S.E. of regression	0.975472	Akaike info criterion	2.961332
Sum squared resid	247.4019	Schwarz criterion	3.698149
Log likelihood	-415.2551	Hannan-Quinn criter.	3.255461
F-statistic	7.353156	Durbin-Watson stat	2.038729
Prob(F-statistic)	0.000000		

Observations

- Prob(F-stat) $= 0$
- $\overline{R^2} = 0.55$
- Many insignificant variables

Table of contents

F-Test : Does the model makes sense ?

- Linear regression : $Y = X\beta + u$
- F-statistic $= 7.50 > F_{0,01}(71; 323 - 71) = 1.55$

$\longrightarrow$ We do reject the null ($\beta = 0$)

- Our model as a whole is statistically meaningful

Analysis of the model

Observations

- Many insignificant variables : too many of them ?
- As expected Competition has a positive effect ($\widehat{\beta} = 0.86$)
- Relationship has a negative effect ($\widehat{\beta} = -0.17$)
- $\overline{R^2} = 0.55$: seems to be a decent model

How to improve the model ?

- Reduce the number of (insignificant) variables
- Try to improve $\overline{R^2}$, to have a better explanation of how often people play sports

Table of contents

The equation

The linear equation of the model

$$\widehat{FREQUENCY} = \widehat{\beta_0} + \widehat{\beta_1} SEG\_SPORT + \widehat{\beta_2} COMPETITION + \widehat{\beta_3} CONT\_BODYB + \widehat{\beta_4} CONT\_PROG + \widehat{\beta_5} MAN + \widehat{\beta_6} LEISURE + \widehat{\beta_7}(M\_SP = 0) + \widehat{\beta_8}(M\_SP = 2) + \widehat{\beta_9}(M\_SP = 4) + \widehat{\beta_{10}}(M\_SP = 8) + \widehat{\beta_{11}}(M\_SP = 10) + \widehat{\beta_{12}}(RAISED = 1) + \widehat{\beta_{13}}(RAISED = 2) + \widehat{\beta_{14}} SP\_DIET + \widehat{\beta_{15}}(TRANSPORT = 1) + \widehat{\beta_{16}} WEIGHT\_WATCH + \widehat{\beta_{17}} WEIGHT + \widehat{\beta_{18}} WORK + \widehat{\beta_{19}}(OCCUP = 1) + \widehat{\beta_{20}}(OCCUP = 2) + \widehat{\beta_{21}}(OCCUP = 3) + \widehat{\beta_{22}}(OCCUP = 4) + \widehat{\beta_{23}}(OCCUP = 9)$$

The linear regression

Observations

- $\overline{R^2} = 0.56$
- Prob(F-Stat) $= 0.0$
- Only 4 insignificant variables
- A total of k $= 13$ variables

Dependent Variable: FREQUENCY
Method: Least Squares
Date: 04/06/16 Time: 13:12
Sample: 1 400
Included observations: 324

	Coefficient	Std. Error	t-Statistic	Prob.
C	2.875668	0.541701	5.308591	0.0000
_SEG_SPORT	0.118774	0.033039	3.595008	0.0004
COMPETITION	0.980277	0.135038	7.259256	0.0000
CONTINUE_BODYBUILD	0.410064	0.126266	3.247624	0.0013
CONTINUE_PROGRESS	0.443956	0.122933	3.611372	0.0004
LEISURE_TIME	-0.002606	0.005283	-0.493295	0.6222
MAN	0.090118	0.139897	0.644170	0.5200
MAIN_SP=0	-1.650757	0.328733	-5.021575	0.0000
MAIN_SP=2	-0.235118	0.161265	-1.457958	0.1459
MAIN_SP=4	0.953365	0.164920	5.780771	0.0000
MAIN_SP=8	0.982023	0.227957	4.307937	0.0000
MAIN_SP=10	-1.441042	0.571206	-2.522808	0.0122
OCCUPATION=1	-0.806955	0.276570	-2.917727	0.0038
OCCUPATION=2	-0.880967	0.439238	-2.005670	0.0458
OCCUPATION=3	-0.957499	0.303999	-3.149674	0.0018
OCCUPATION=4	-0.944708	0.311981	-3.028095	0.0027
OCCUPATION=9	-0.898988	0.371104	-2.422466	0.0160
RAISED=1	-0.426458	0.164229	-2.596718	0.0099
RAISED=2	-0.305022	0.164740	-1.851535	0.0651
SPORT_DIET	0.547719	0.137337	3.988147	0.0001
TRANSPORT=1	-0.173781	0.129852	-1.338293	0.1818
WEIGHT_WATCH	-0.279486	0.118667	-2.355209	0.0192
WEIGHT	0.005548	0.005682	0.976407	0.3296
WORK	-0.011906	0.005470	-2.176490	0.0303

R-squared	0.592857	Mean dependent var	3.043210
Adjusted R-squared	0.561642	S.D. dependent var	1.448171
S.E. of regression	0.958813	Akaike info criterion	2.824947
Sum squared resid	275.7969	Schwarz criterion	3.105002
Log likelihood	-433.6413	Hannan-Quinn criter.	2.936729
F-statistic	18.99309	Durbin-Watson stat	1.967294
Prob(F-statistic)	0.000000		

Figure 5: Our survey.

The coefficiented equation

The linear equation of the model

$$\widehat{FREQUENCY} = 2.88 + 0.12 * SEG\_SPORT +$$
$$0.98 * COMPETITION + 0.41 * CONT\_BODYB +$$
$$0.44 * CONT\_PROG + 0.09 * MAN - 0.003 * LEISURE +$$
$$1.65 * (M\_SP = 0) + 0.23 * (M\_SP = 2) + 0.95 * (M\_SP = 4) +$$
$$0.98 * (M\_SP = 8) - 1.44 * (M\_SP = 10) - 0.43 * (RAISED = 1)$$
$$-0.31 * (RAISED = 2) + 0.55 * SP\_DIET - 0.17(TRANSPORT = 1)$$
$$-0.28 * WEIGHT\_WATCH + 0.005 * WEIGHT - 0.01 * WORK$$
$$-0.80 * (OCCUP = 1) - 0.88 * (OCCUP = 2) - 0.96 * (OCCUP = 3)$$
$$-0.95 * (OCCUP = 4) - 0.89 * (OCCUP = 9)$$

Table of contents

Analysis of the second model

Observations

- $\overline{R^2} = 0.56$ & Prob(F-stat) $= 0.0$:
  $\rightarrow$ No precision lost despite removal of many variables
- Only 4 insignificant variables : great improvement !
- Some marginal effects seem odd ($\widehat{\beta}_{racket\ sports} < 0$)

The variables and the marginal effects (1)

Variable	Expected effect	Real effect	Significant
# of secondary sports	+	+	Yes
Competition	+	+	Yes
Continue_bodybuild	+	+	Yes
Continu_progress	+	+	Yes
Leisure time	−	−	No
Man	+	+	No
main_sp = 0 (No Sport)	−	−	Yes
main_sp = 2 (Racket sports)	+	−	Yes
main_sp = 4 (Swimming)	+	+	Yes
main_sp = 8 (Body building)	+	+	Yes
main_sp = 10 (Martial arts)	+	−	Yes
Sport_diet	+	+	Yes

Table 2: Comparison of expected and real marginal effects (1)

The variables and the marginal effects (2)

Variable	Expected effect	Real effect	Significant
Raised = 1 (Countryside)	−	−	Yes
Raised = 2 (Small city)	+	−	Yes
Transport = 1 (By foot)	−	−	No
Weight_watch	+	−	Yes
Weight	?	+	No
Work	−	−	Yes
Occupation = 1 (Student)	+	−	Yes
Occupation = 2 (No job)	?	−	Yes
Occupation = 3 (Executive)	−	−	Yes
Occupation = 4 (Employee)	−	−	Yes
Occupation = 9 (Intermediate job)	?	−	Yes

Table 3: Comparison of expected and real marginal effects (2)

Table of contents

Normality of the errors

Observations

- Mean ≈ 0
- $JB \sim \chi^2(2)$
- $JB_{corrected} = \frac{n-k}{n} * JB_{Eviews}$
- $JB_{corrected} = 1.45 < \chi^2_{0.95}(2) = 5.99$
- Normality of errors assumed

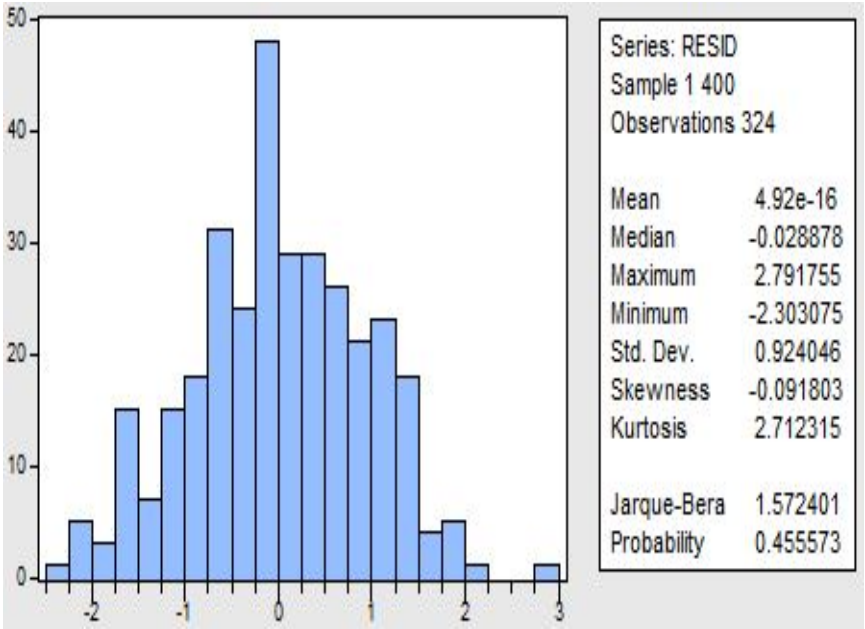

Figure 6: Graph of residuals

What about heteroscedasticity ?

Heteroskedasticity Test: White

F-statistic	1.117807	Prob. F(216,107)	0.2605
Obs*R-squared	224.5069	Prob. Chi-Square(216)	0.3314
Scaled explained SS	164.7919	Prob. Chi-Square(216)	0.9961

Figure 7: White's test of heteroscedasticity

- $nR^2 = 224.51 < \chi^2_{0.95}(2) = 255.26$
- We do not reject the null
- We assume homoscedasticity : $Var(u) = \sigma^2 In$

What about linearity ?

Ramsey RESET Test:

F-statistic	0.706010	Prob. F(1,299)	0.4014
Log likelihood ratio	0.764139	Prob. Chi-Square(1)	0.3820

Figure 8: Ramsey test of linearity

- We do not reject the null
- We assume linearity of the model : $Y = X\beta + u$

Table of contents

Pros and cons

- We obtained a meaningful model
- ALL the classical assumptions hold

- FREQUENCY based on ranges rather than precise figures
- $\overline{R^2}$ is decent but some useful explanatory variables might have been forgotten in the survey

Conclusion : possible improvements ?

- A satisfying model but far from being perfect

Limits

- Fairly homogeneous population
- Data sometimes lacked of precision : ranges in the survey
- Other dependent variables could have been used