What is your student performance

Alexandre Lohéac, Emilien Verdier

ENAC

April 23, 2019

What our goal is

- Find a reasonable model of absences based on acquired data
- Analyze its limitations and its validity
- Interpret the result and compare it to expectations
- Potential uses of this model for schools

Introduction to the data

- $n = 349$
- $k = 30$
- All students from 15 to 22 years old in high school
- Mathematics

Data collected during the 2005- 2006 school year from two public schools, from the Alentejo region of Portugal.

Variables used

Name	Variable	Type	Explanation
Sex	sex	binary	0-Female, 1-Male
Age	age	numeric	Ranging from 15 to 22 included
Address	adress	category	1-Rural area, 2-Urban area
Family size	famsize	category	1-Less than 3, 2-Greater than 3
Parent status	pstatus	category	Cohabitation: 1-together, 2-apart
Mother education	medu	numeric	Level of studies (1 to 4)
Father education	fedu	numeric	Level of studies (1 to 4)
Reason	reason	category	1-Home 2-Reputation 3-Course 4-Other
Guardian	guardian	category	0-Female, 1-Male, 2-Other
Travel time	traveltime	category	1-3, 15 minute
Studytime	studytime	category	1-2h, 4-10h
Failures	failures	numeric	Classes failed in the past
Support	schoolsup	binary	Extra educational support

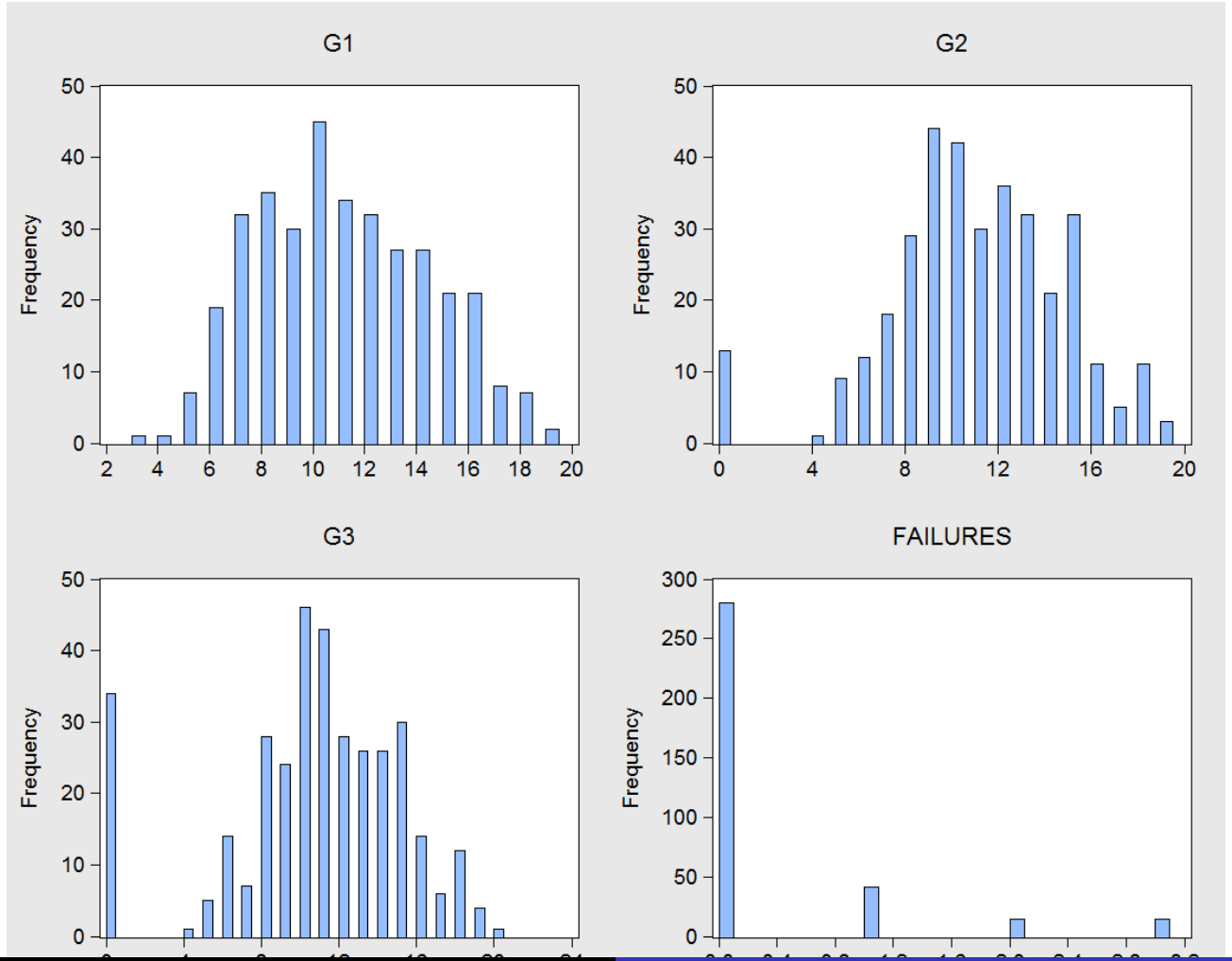
Variables used

Name	Variable	Type	Explanation
Family support	famsup	binary	family educational support
Paid class	paid	binary	Extra paid class in the course
Activities	activities	binary	Has afterschool activities
Nursery	nursery	binary	Attended nursury school
Higher education	higher	binary	Wants to further studies
Internet	internet	binary	Has access to the Internet
Relationship	romantic	binary	Is in a romantic relationship
Family relationship	famrel	numeric	1-Very bad, 5-Excellent
Free time	freetime	numeric	1-Very low, 5-Very high
Going out	goout	numeric	1-Never, 5-Often
Alcohol consumption	dalc	numeric	Workday: 1-low, 5-high
Alcohol consumption	walc	numeric	Weekend: 1-low, 5-high
Health	health	numeric	Status: 1-very bad, 5-very good

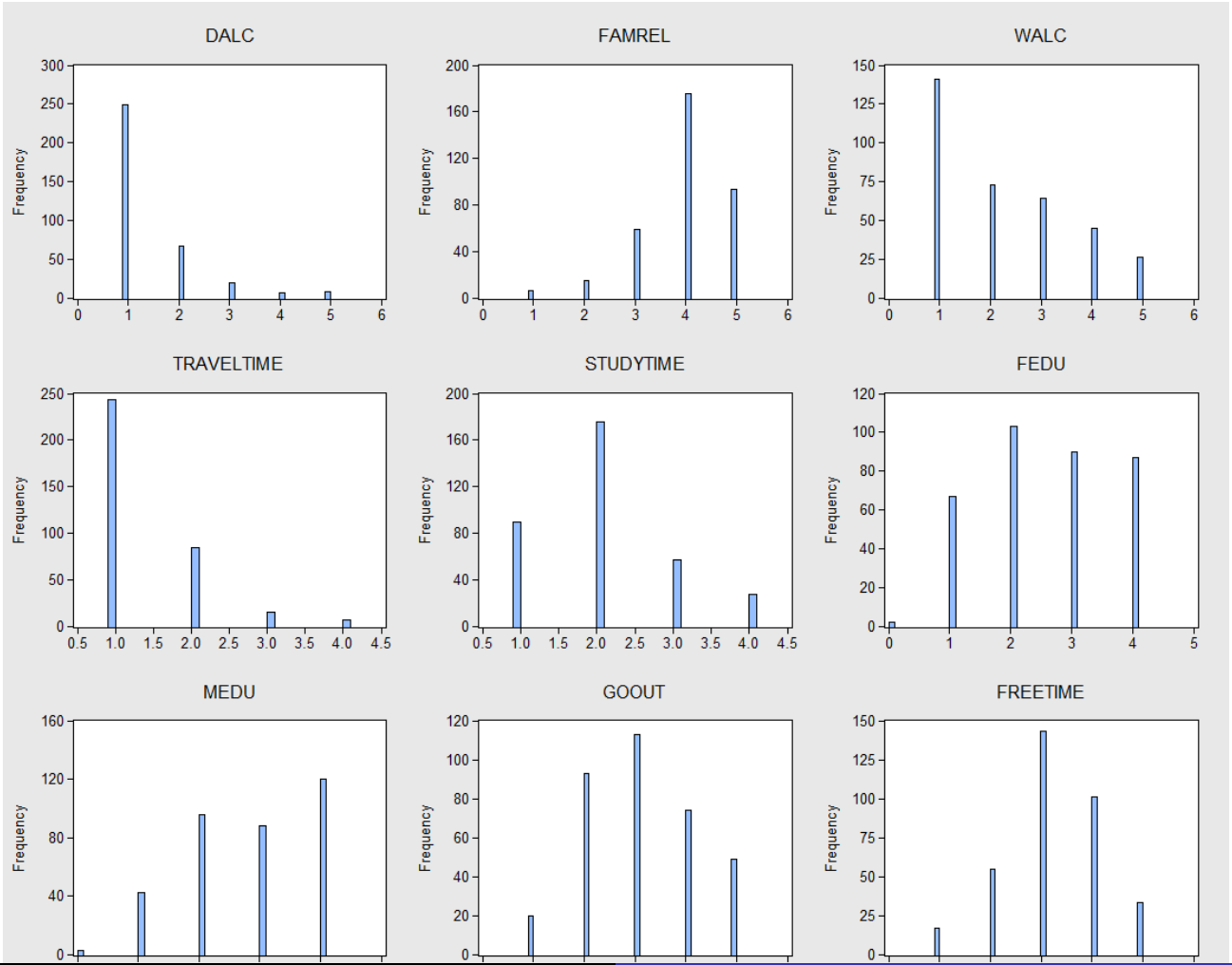
Variables used

Name	Variable	Type	Explanation
Absences	absences	numeric	number of school absences (0-93)
Grade 1	g1	numeric	Grade first semester (0-20)
Grade 2	g2	numeric	Grade second semester (0-20)
Grade 3	g3	numeric	Total grade (0-20)

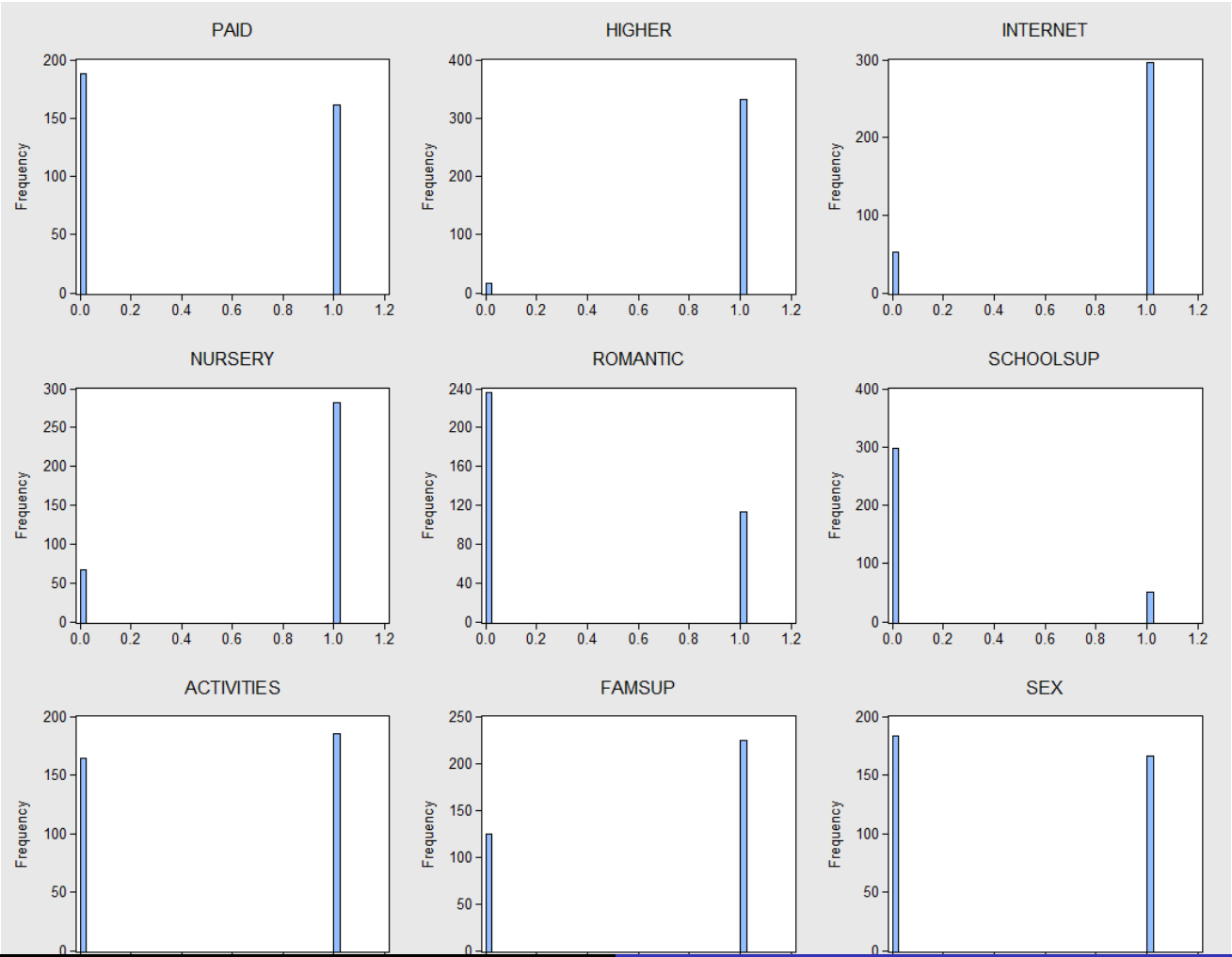
Data overview



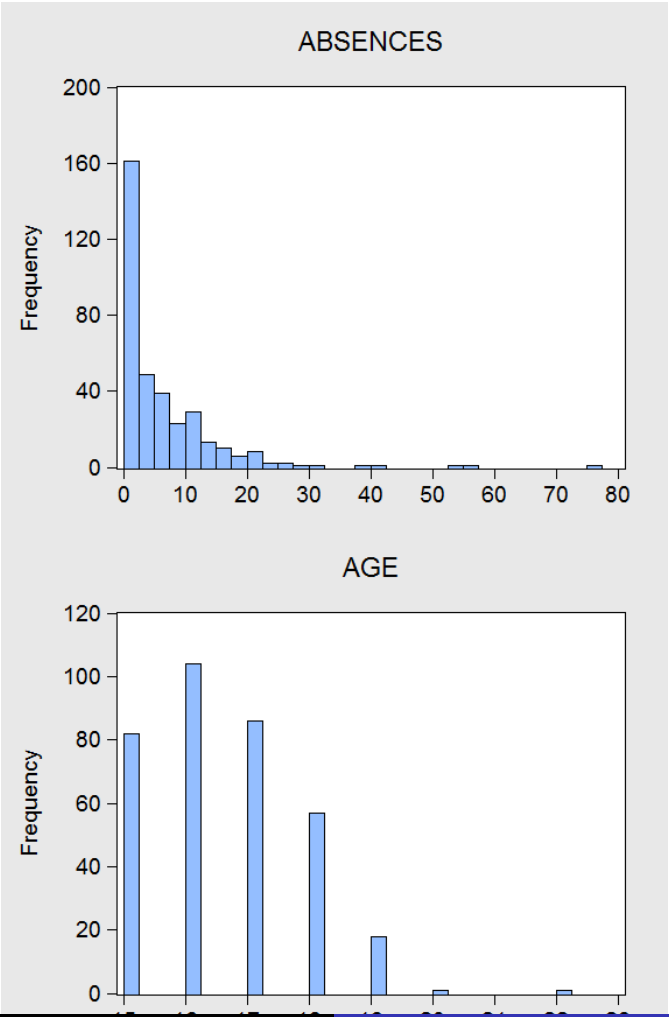
Data overview



Data overview



Data overview



The first mathematical model

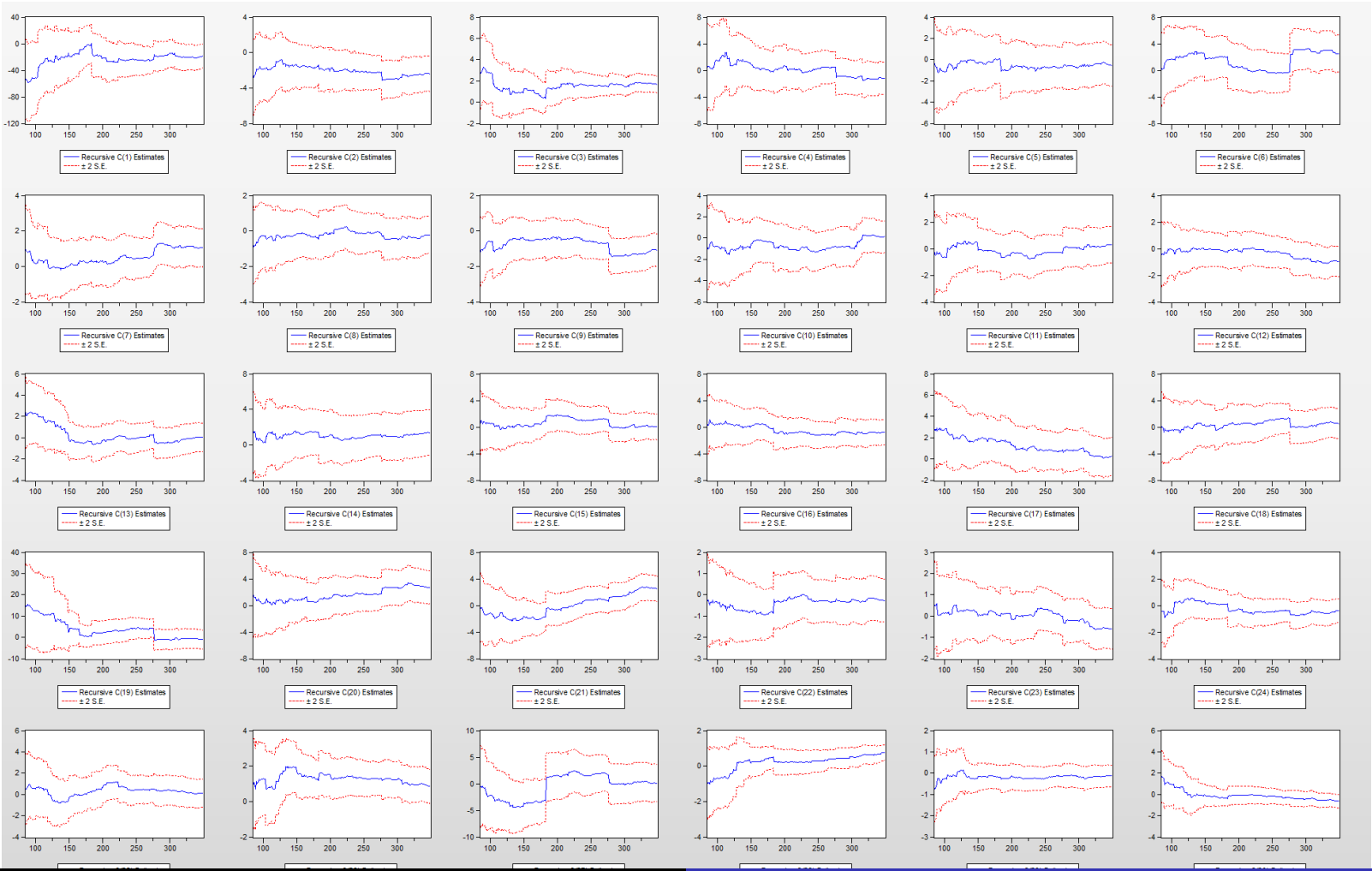
Dependent Variable: ABSENCES
Method: Least Squares
Date: 04/12/19 Time: 15:29
Sample: 1 349
Included observations: 349

	Coefficient	Std. Error	t-Statistic	Prob.
C	-17.96303	9.168090	-1.959300	0.0509
SEX	-2.368637	0.980276	-2.416295	0.0162
AGE	1.673079	0.397359	4.210493	0.0000
ADDRESS	-1.252369	1.195253	-1.047786	0.2955
FAMSIZE	-0.601219	0.964909	-0.623083	0.5337
PSTATUS	2.552975	1.386353	1.841505	0.0665
MEDU	1.024129	0.537524	1.905272	0.0576
FEDU	-0.221204	0.521404	-0.424247	0.6717
REASON	-1.090163	0.457134	-2.384775	0.0177
GUARDIAN	0.173488	0.738971	0.234769	0.8145
TRAVELTIME	0.307675	0.684953	0.449192	0.6536
STUDYTIME	-1.018217	0.557393	-1.826749	0.0687
SCHOOLSUP	1.383851	1.294890	1.068702	0.2860
FAMSUP	-0.020665	0.959650	-0.021533	0.9828
PAID	-0.803658	0.927335	-0.866632	0.3868
ACTIVITIES	0.216825	0.878705	0.246755	0.8053
NURSERY	0.512331	1.123446	0.456035	0.6487
HIGHER	-1.121554	2.191567	-0.511759	0.6092
INTERNET	2.750375	1.244650	2.209758	0.0278
ROMANTIC	2.378679	0.939125	2.532867	0.0118
FAMREL	-0.317976	0.506387	-0.627931	0.5305
FREETIME	-0.646353	0.469210	-1.377534	0.1693
GOOUT	-0.386254	0.452398	-0.853792	0.3939
DALC	0.184079	0.662844	0.277710	0.7814
WALC	0.805845	0.479130	1.681892	0.0936
HEALTH	0.056708	0.314757	0.180165	0.8571
G3	0.749093	0.224208	3.341062	0.0009
G1	-0.120304	0.255286	-0.471250	0.6378
G2	-0.653642	0.303171	-2.156021	0.0318

R-squared	0.210382	Mean dependent var	5.965616
Adjusted R-squared	0.141291	S.D. dependent var	8.341764
S.E. of regression	7.730025	Akaike info criterion	7.007540
Sum squared resid	19121.05	Schwarz criterion	7.327875
Log likelihood	-1193.816	Hannan-Quinn criter.	7.135058

Variable	Expectation	Model	Significant
SEX	+	-	yes
AGE	+	+	yes
PSTATUS	-	+	yes
TRAVELTIME	+	+	no
ROMANTIC	?	+	yes
HIGHER	-	-	no
INTERNET	+	+	yes
FAMREL	+	-	no
GOOUT	+	-	no
DALC	+	+	no
WALC	+	+	yes
HEALTH	+	+	no

Recursive calculations



Recursive calculations

- High changes in some places,
- Peak at 277, but no major changes in expl. variables
- => Missing variables

obs	Actual	Fitted	Residual	Residual Plot
271	15.0000	11.9465	3.05349	
272	4.00000	6.44752	-2.44752	
273	2.00000	3.80065	-1.80065	
274	2.00000	3.53623	-1.53623	
275	2.00000	7.25263	-5.25263	
276	6.00000	9.13382	-3.13382	
277	75.0000	18.4664	56.5336	
278	22.0000	11.4470	10.5530	
279	15.0000	12.6272	2.37279	
280	8.00000	7.97045	0.02955	
281	30.0000	13.9028	16.0972	
282	19.0000	11.9492	7.05082	
283	1.00000	4.02528	-3.02528	
284	4.00000	9.06681	-5.06681	
285	4.00000	4.56257	-0.56257	

Endogeneity

- FREETIME correlated with ACTIVITES
- ABSENCES with G1, G2, G3 and
- FAILURES $=>$ G1, G2, G3
- WALC,DALC

An improved model

DALC	-63.29662	50.88308	-1.243962	0.2182
WALC	-6.064613	30.65386	-0.197842	0.8438
1/HEALTH	-2.079078	3.249946	-0.639727	0.5247
G3	-0.878793	0.897535	-0.979118	0.3313
G1	-0.153715	0.623536	-0.246521	0.8061
G2	1.148052	1.060027	1.083040	0.2830
WALC*AGE	0.487519	1.965494	0.248039	0.8049
DALC*AGE	4.106733	3.275914	1.253614	0.2147
PAID*AGE	1.446959	3.347584	0.432240	0.6671
ACTIVITIES*FREETIME	-2.621356	1.854134	-1.413790	0.1624
GOOUT*FREETIME	-0.734580	0.869661	-0.844673	0.4015
FAMSUP*FAMREL	1.350704	2.577756	0.523984	0.6022

R-squared	0.359476	Mean dependent var	4.760000
Adjusted R-squared	-0.022772	S.D. dependent var	6.781019
S.E. of regression	6.857794	Akaike info criterion	6.970613
Sum squared resid	2915.819	Schwarz criterion	7.960577
Log likelihood	-310.5306	Hannan-Quinn criter.	7.371269
F-statistic	0.940426	Durbin-Watson stat	1.773381
Prob(F-statistic)	0.572456		

- Only sample of first 100
- Included Age+Age^2
- Better R^2
- 1/Health

A binary model

PAID	0.134365	2.562007	0.052445	0.9582
ACTIVITIES	0.289863	0.607663	0.477013	0.6334
NURSERY	-0.033968	0.218696	-0.155322	0.8766
HIGHER	-0.450956	0.444058	-1.015532	0.3099
INTERNET	-0.106297	0.243043	-0.437361	0.6618
ROMANTIC	0.010465	0.187389	0.055847	0.9555
FAMREL	0.467376	0.184415	2.534372	0.0113
FREETIME	0.330698	0.237746	1.390974	0.1642
GOOUT	-0.183271	0.277359	-0.660773	0.5088
DALC	-0.793433	2.006334	-0.395464	0.6925
WALC	-0.809909	1.457434	-0.555709	0.5784
1/HEALTH	-0.311299	0.363052	-0.857450	0.3912
G3	-0.344635	0.065387	-5.270704	0.0000
G1	0.095872	0.051912	1.846820	0.0648
G2	0.225076	0.074792	3.009345	0.0026
WALC*AGE	0.043606	0.088474	0.492860	0.6221
DALC*AGE	0.046374	0.119966	0.386562	0.6991
PAID*AGE	-0.009944	0.156568	-0.063514	0.9494
ACTIVITIES*FREETIME	-0.207120	0.177259	-1.168462	0.2426
GOOUT*FREETIME	-0.013439	0.075761	-0.177381	0.8592
FAMSUP*FAMREL	-0.151544	0.226281	-0.669718	0.5030

McFadden R-squared	0.239317	Mean dependent var	0.292264
S.D. dependent var	0.455456	S.E. of regression	0.408915
Akaike info criterion	1.136922	Sum squared resid	52.00265
Schwarz criterion	1.556672	Log likelihood	-160.3929
Hannan-Quinn criter.	1.304015	Deviance	320.7859
Restr. deviance	421.7079	Restr. log likelihood	-210.8540
LR statistic	100.9220	Avg. log likelihood	-0.459579

- P(absences <= 1)
- based on second model

Expectation-Prediction Evaluation for Binary Specification
Equation: EQUBIN
Date: 04/15/19 Time: 14:02
Success cutoff: C = 0.5

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	232	55	287	247	102	349
P(Dep=1)>C	15	47	62	0	0	0
Total	247	102	349	247	102	349
Correct	232	47	279	247	0	247
% Correct	93.93	46.08	79.94	100.00	0.00	70.77
% Incorrect	6.07	53.92	20.06	0.00	100.00	29.23
Total Gain*	-6.07	46.08	9.17			
Percent Gain**	NA	46.08	31.37			

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
E(# of Dep=0)	194.32	51.82	246.14	174.81	72.19	247.00
E(# of Dep=1)	52.68	50.18	102.86	72.19	29.81	102.00
Total	247.00	102.00	349.00	247.00	102.00	349.00
Correct	194.32	50.18	244.50	174.81	29.81	204.62
% Correct	78.67	49.20	70.06	70.77	29.23	58.63
% Incorrect	21.33	50.80	29.94	29.23	70.77	41.37
Total Gain*	7.90	19.97	11.43			
Percent Gain**	27.02	28.22	27.62			

*Change in "% Correct" from default (constant probability) specification

Conclusion

Conclusion