Etude d'une fonction de salaire

Nous disposons d'une enquête américaine, datant de 1985, sur l'emploi de 258 personnes ayant fini leurs études. Pour chaque individu, nous avons des informations sur son salaire horaire (en logarithme), noté lsal, son niveau d'éducation, educ, son sexe (1 si femme et 0 si homme), noté sexe, son expérience professionnelle en années, notée expro.

La régression de lsal sur les variables educ, sexe et expro donne les résultats suivants :

```
proc reg data=a.sal;
model lsal=educ sexe expro;
output out=a.res p=ajust r=res;
run;quit;
```

The REG Procedure
Model: MODEL1
Dependent Variable: lsal

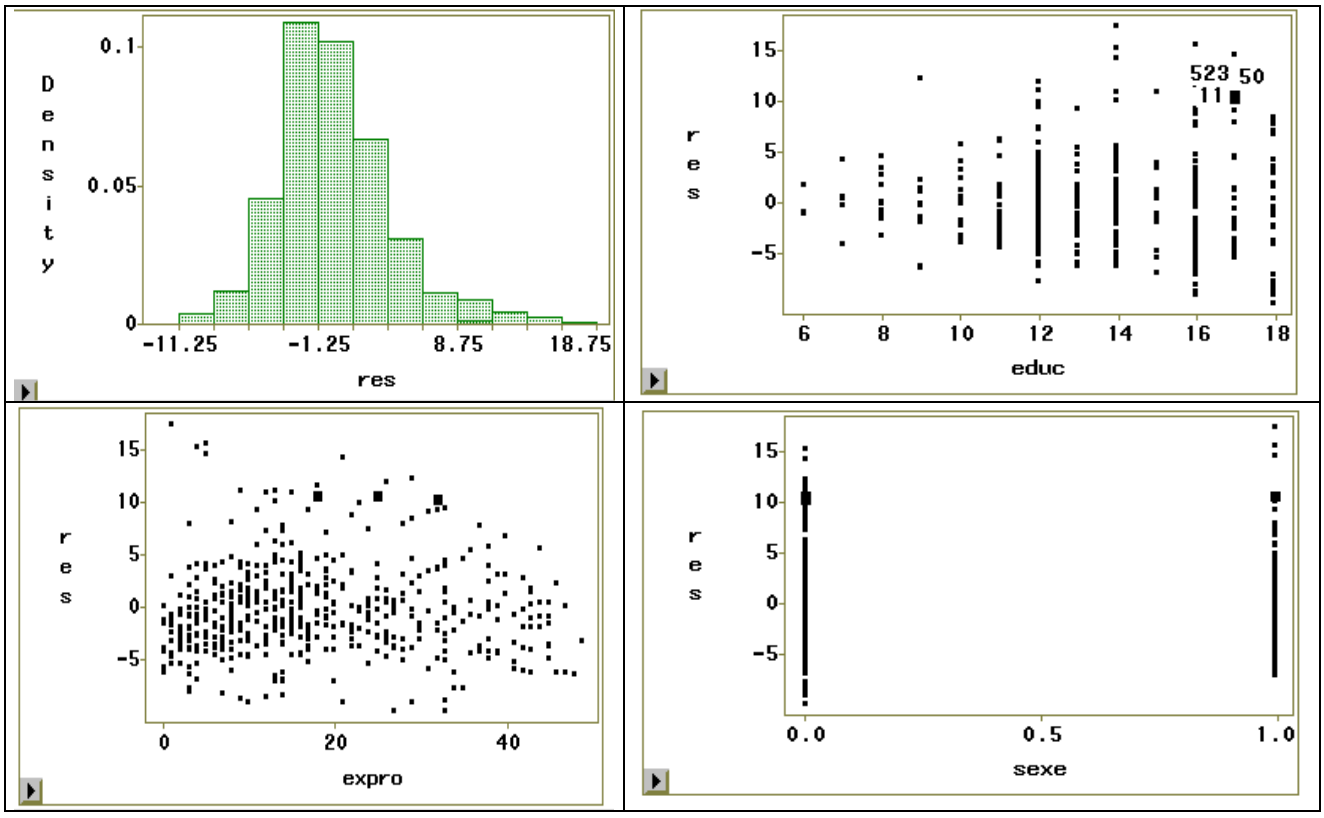
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3703.52999	1234.51000	70.14	<.0001
Error	524	9222.86615	17.60089		
Corrected Total	527	12926			

Root MSE	4.19534	R-Square	0.2865
Dependent Mean	9.00966	Adj R-Sq	0.2824
Coeff Var	46.56493		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4.77392	1.15947	-4.12	<.0001
educ	1	0.97962	0.07753	12.64	<.0001
sexe	1	-2.40796	0.36789	-6.55	<.0001
expro	1	0.11755	0.01594	7.38	<.0001



Afin de tester l'hypothèse d'homoscédasticité par le test de White, on estime la régression du carré des résidus de l'estimation précédente sur un ensemble de variables :

```
data a.het;
set a.res;
res2=res*res;
educ2=educ*educ; sexe2=sexe*sexe; expro2=expro*expro;
educsexe=educ*sexe; educexpro=educ*expro; sexeexpro=sexe*expro;
run;
proc reg data=a.het;
model res2=educ sexe expro educ2 expro2 educsexe educexpro sexeexpro;run;
quit;
```

The REG Procedure
Model: MODEL1
Dependent Variable: res2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	57249	7156.09182	6.71	<.0001
Error	519	553647	1066.75723		
Corrected Total	527	610896			

Root MSE	32.66125	R-Square	0.0937
Dependent Mean	17.46755	Adj R-Sq	0.0797
Coeff Var	186.98244		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.81311	45.93924	0.30	0.7638
educ	1	-3.34478	6.01094	-0.56	0.5781
sexe	1	16.81911	18.27653	0.92	0.3579
expro	1	-0.50104	0.98542	-0.51	0.6113
educ2	1	0.26070	0.19897	1.31	0.1907
expro2	1	0.00915	0.01026	0.89	0.3727
educsexe	1	-1.33710	1.22890	-1.09	0.2771
educexpro	1	0.04359	0.05769	0.76	0.4503
sexeexpro	1	-0.40888	0.25100	-1.63	0.1039

The REG Procedure
Model: MODEL2
Dependent Variable: res2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	38906	38906	35.78	<.0001
Error	526	571990	1087.43316		
Corrected Total	527	610896			

Root MSE	32.97625	R-Square	0.0637
Dependent Mean	17.46755	Adj R-Sq	0.0619
Coeff Var	188.78579		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-27.70184	7.68673	-3.60	0.0003
educ	1	3.45144	0.57702	5.98	<.0001

Afin d'effectuer le test de Goldfeld-Quandt, on effectue les régressions suivantes

```
proc sort data=a.sal;by educ;run;
data educ1;set a.sal;if i<177;run;
data educ2;set a.sal;if i>352;run;

proc reg data=educ1;
educ1: model lsal=educ sexe expro;run;quit;
proc reg data=educ2;
educ3: model lsal=educ sexe expro;run;quit;
```

The REG Procedure
Model: educ1
Dependent Variable: lsal

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1652.31358	550.77119	29.99	<.0001
Error	172	3159.03690	18.36649		
Corrected Total	175	4811.35049			

Root MSE	4.28561	R-Square	0.3434
Dependent Mean	9.69153	Adj R-Sq	0.3320
Coeff Var	44.22019		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.44848	2.03697	-3.17	0.0018
educ	1	1.14998	0.13511	8.51	<.0001
sexe	1	-2.20881	0.65125	-3.39	0.0009
expro	1	0.10755	0.02783	3.86	0.0002

The REG Procedure
Model: educ3
Dependent Variable: lsal

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	882.38505	294.12835	18.51	<.0001
Error	172	2732.77540	15.88823		
Corrected Total	175	3615.16045			

Root MSE	3.98600	R-Square	0.2441
Dependent Mean	8.68091	Adj R-Sq	0.2309
Coeff Var	45.91690		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.39845	1.97406	-1.72	0.0869
educ	1	0.86353	0.13699	6.30	<.0001
sexe	1	-2.26074	0.60936	-3.71	0.0003
expro	1	0.11664	0.02889	4.04	<.0001

```
data null;
fi=18.36649/15.88823; pval=1-cdf('F',fi,172,172); run;
```

f	pval
1.15598	0.17139

Finalement, on fait une estimation de type Moindres Carrés Pondérés :

```
data a.mcp;
set a.sal;
c=1;
lsalseduc=lsal/sqrt(educ);
educseduc=educ/sqrt(educ);
sexeseduc=sexe/sqrt(educ);
exproseduc=expro/sqrt(educ);
cseduc=c/sqrt(educ);
run;

proc reg data=a.mcp;
MCP: model lsalseduc=cseduc educseduc sexeseduc exproseduc/noint;run;
quit;
```

The REG Procedure
Model: MCP
Dependent Variable: lsalseduc

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3394.57828	848.64457	666.68	<.0001
Error	524	667.02048	1.27294		
Uncorrected Total	528	4061.59875			

Root MSE	1.12825	R-Square	0.8358
Dependent Mean	2.47340	Adj R-Sq	0.8345
Coeff Var	45.61524		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
cseduc	1	-4.32784	1.09903	-3.94	<.0001
educseduc	1	0.95557	0.07479	12.78	<.0001
sexeseduc	1	-2.42027	0.35090	-6.90	<.0001
exproseduc	1	0.11044	0.01499	7.37	<.0001