ENSAE-ENSAI Formation continue [Cepe]

Certificat de
chargé d'études statistiques

Sessions de révision – Exercices pratiques

Martin Chevalier (INSEE)

La plupart des exercices pratiques proposés dans le cadre des sessions de révision des 2, 3 et 4 mai s'appuie sur des exploitations de l'enquête PISA (Program for International Student Assessment) 2012. Réalisée tous les trois ans par l'OCDE dans une soixantaine de pays, cette enquête vise à mesurer les acquis des élèves de 15 ans.

En plus des scores aux tests standardisés de mathématiques, compréhension de l'écrit et sciences, cette enquête comporte de très nombreuses informations sur l'origine sociale des élèves, leurs conditions d'enseignement ainsi que leur rapport aux enseignants et à l'école.

Du point de vue de la formation, cette enquête présente ainsi l'avantage de comporter une très large variété de variables qualitatives et quantitatives. Elle se prête ainsi à tous les outils et méthodes au programme des sessions de révision des 2, 3 et 4 mai.

Les fichiers de l'enquête PISA 2012 sont librement téléchargeables sur le site de l'OCDE[1]*. Le fichier « élèves » réduit*[2] *ainsi que le code ayant servis à la production des sorties statistiques utilisées dans les exercices pratiques sont fournis aux stagiaires.*

1. `http://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm`

2. L'échantillon est restreint aux données collectées en France, rééchantillonnées à hauteur de 30 % pour limiter la taille du fichier final.

Session 1 : Statistique descriptive

Question 1.1 La variable `ST01Q01` correspond à la classe dans laquelle se trouve l'élève au moment de l'enquête (la $10^{\text{ème}}$ classe correspond à la seconde en France). L'enquête permet d'obtenir les deux tris à plat suivants :

ST01Q01	Frequency	Percent	Cumulative Frequency	Cumulative Percent
8	22	1.61	22	1.61
9	386	28.22	408	29.82
10	910	66.52	1318	96.35
11	49	3.58	1367	99.93
12	1	0.07	1368	100.00

ST01Q01	Frequency	Percent	Cumulative Frequency	Cumulative Percent
8	3441.067	1.67	3441.067	1.67
9	59428.85	28.79	62869.91	30.45
10	135518.9	65.64	198388.8	96.10
11	7890.986	3.82	206279.8	99.92
12	165.3195	0.08	206445.1	100.00

a. Quelle procédure a été utilisée pour produire ces résultats ?

b. Interprétez les résultats du tri à plat de gauche. Comment recoderiez-vous cette variable pour en mener l'étude ?

c. Comparez les résultats des deux tris à plat. À quoi les différences observées sont-elles dues à votre avis ? Quel tri à plat privilégieriez-vous et pourquoi ?

Question 1.2 La variable `PV1MATH` correspond au score synthétique de l'élève aux évaluations de mathématique. Le tableau suivant en synthétise la distribution (non-pondérée) :

Analysis Variable : PV1MATH					
N	**Mean**	**Median**	**Variance**	**Minimum**	**Maximum**
1368	498.1789377	498.6836000	9232.63	230.8070000	781.4379000

a. Quelle procédure a été utilisée pour produire ces résultats ?

b. Interprétez la valeur de la moyenne. En analysant les autres statistiques présentées, que pensez-vous de l'influence des valeurs extrêmes ?

c. Calculez l'écart-type et le coefficient de variation de la variable `PV1MATH`.

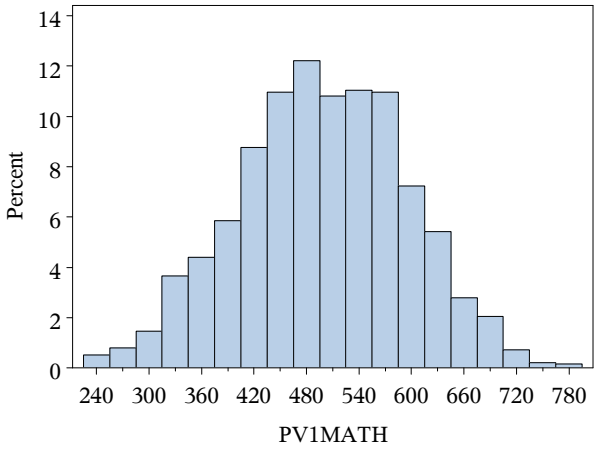
Question 1.3 Les deux graphiques suivants représentent la distribution de la variable `PV1MATH`.

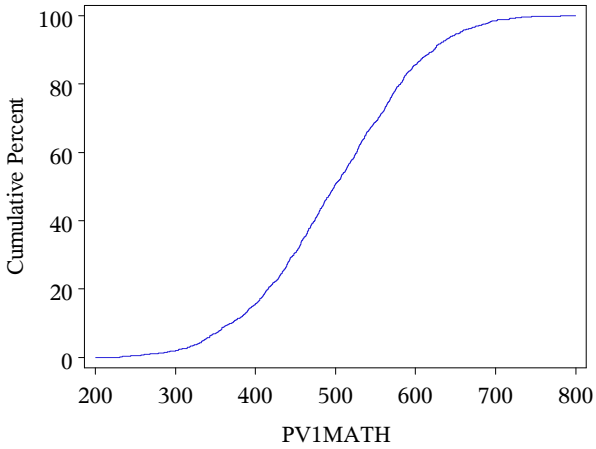
a. Quel est le nom de ces deux graphiques ? En quoi leur analyse confirme-t-elle les résultats de la question précédente quant à l'influence des valeurs extrêmes ?

b. Utilisez ces graphiques pour déterminer (approximativement) la valeur des quartiles de la variable `PV1MATH` ainsi que celle des premier et neuvième déciles.

c. Quelle procédure auriez-vous pu utiliser pour obtenir directement les quantiles de la variable `PV1MATH` [3] ?

3. C'est en fait cette même procédure qui a été utilisée pour produire ces graphiques.





Question 1.4 La variable `ST01Q01` est recodée en deux modalités dans la variable `retard` qui vaut 1 si l'élève est « en retard » (s'il est dans une classe inférieure à la seconde au moment de l'enquête) et 0 sinon. Cette variable est croisée avec la variable `ST04Q01`, qui code le sexe des élèves.

Frequency Expected Cell Chi-Square Percent Row Pct Col Pct

Table of ST04Q01 by retard			
	retard		
ST04Q01	0	1	Total
Femme	528 501.05 1.4493 38.60 73.95 55.00	186 212.95 3.41 13.60 26.05 45.59	714 52.19
Homme	432 458.95 1.5822 31.58 66.06 45.00	222 195.05 3.7229 16.23 33.94 54.41	654 47.81
Total	960 70.18	408 29.82	1368 100.00

Remarque : La modalité `"1"` de la variable `ST04Q01` correspond aux femmes et la modalité `"2"` aux hommes, à rebours des conventions françaises. Un label a été appliqué pour plus de clarté.

a. Quelles statistiques déjà présentes dans la question 1 retrouvez-vous dans ce tableau ? Comment les désigne-t-on dans le contexte du tri croisé ?

b. Interprétez un pourcentage de cellule, un pourcentage en ligne et un pourcentage en colonne.

c. Utilisez l'ensemble des informations du tableau pour identifier et justifier des sur- ou sous-représentations manifestes.

Question 1.5 Les variables `PV1READ` et `PV1SCIE` correspondent respectivement aux scores synthétiques de l'élève aux évaluations de compréhension de l'écrit et de sciences. Le

tableau suivant représente le coefficient de corrélation de Pearson calculés entre les trois scores pris deux-à-deux :

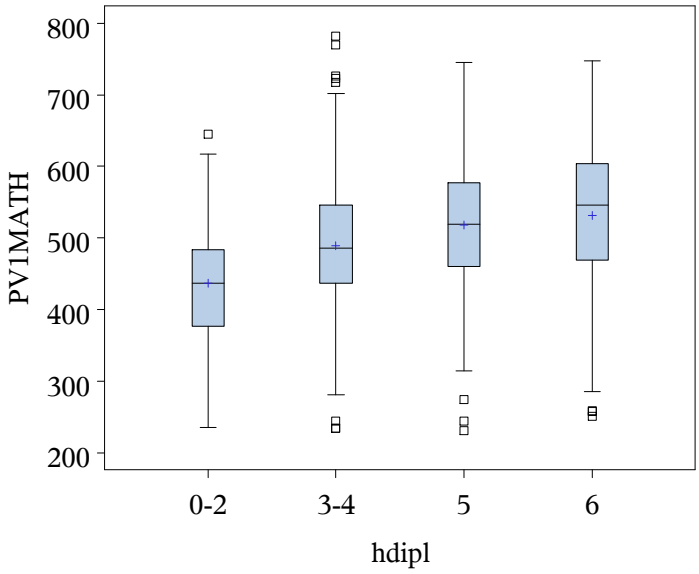
Pearson Correlation Coefficients, N = 1368 Prob > |r| under H0: Rho=0			
	PV1MATH	**PV1READ**	**PV1SCIE**
PV1MATH	1.00000	0.86722 <.0001	0.89773 <.0001
PV1READ	0.86722 <.0001	1.00000	0.88809 <.0001
PV1SCIE	0.89773 <.0001	0.88809 <.0001	1.00000

a. Quelle procédure a été utilisée pour produire ces résultats ?

b. Interprétez la valeur des indicateurs. Que peut-on conclure ?

c. Des indicateurs analogues existent-ils ? Quel est leur intérêt et comment les calculer ?

Question 1.6 On dispose par ailleurs du plus haut niveau d'étude des parents que l'on regroupe en quatre modalités [4] : 0-2 Aucun, primaire ou collège ; 3-4 Fin d'études secondaires ou post-secondaire non-supérieur ; 5 Premier cycle d'études supérieures ; 6 Seconde cycle et au-delà. Le graphique suivant représente la relation entre score en mathématiques (variable `PV1MATH`) et plus haut diplôme des parents regroupé (variable `hdipl`).



a. Quel nom porte ce type de graphique et avec quelle procédure le construire ?

b. Explicitez la signification des éléments constituant une boîte et interprétez le graphique.

c. Quelle mesure d'association correspond à ce type de représentation ? Sa valeur est 0,092048 : qu'en pensez-vous ?

4. La nomenclature originale est ISCED 1997 (`http://www.uis.unesco.org/Library/Documents/isced97-fr.pdf`).

Session 2 : Statistique inférentielle

Question 2.1 À partir des informations suivantes, construisez l'intervalle de confiance à 95 % de la moyenne des scores de mathématiques, compréhension de l'écrit et sciences :

Variable	N	Mean	Std Dev
PV1MATH	1368	498.1789377	96.0865993
PV1READ	1368	509.9585724	108.6969627
PV1SCIE	1368	503.4514466	98.0487839

Quelle procédure permet d'obtenir directement ces intervalles de confiance ?

Question 2.2 On cherche à tester l'hypothèse

$$H_0 : \beta = c \quad \text{contre} \quad H_1 : \beta > c$$

On sait que sous H_0, une statistique Z suit une loi du χ^2 à 8 degrés de liberté.

a. Ce test est-il un test bilatéral ou unilatéral ?

b. On souhaite mener ce test au niveau de confiance de 95 %. Utilisez les quantiles de la loi du χ^2 en annexe pour déterminer la région critique. Quelle est la valeur critique ?

c. Le calcul de Z donne la valeur 17,35. Que concluez-vous ?

d. Afin d'être plus prudent, on préfère en fait ne tolérer un risque de première espèce que de 1 % : cela modifie-t-il votre conclusion ?

Question 2.3 On cherche à tester l'hypothèse

$$H_0 : \beta = c \quad \text{contre} \quad H_1 : \beta \neq c$$

On sait que sous H_0, une statistique Z suit une loi de Student à 4 degrés de liberté.

a. Ce test est-il un test bilatéral ou unilatéral ?

b. On souhaite mener ce test au niveau de confiance de 95 %. Utilisez les quantiles de la loi de Student en annexe pour déterminer la région critique correspondante. Est-il possible de l'écrire sous la forme $[q; +\infty[$?

c. Le calcul de Z donne la valeur 7,35. Que concluez-vous ?

d. Votre conclusion serait-elle modifiée si le test était mené à 99 % ?

Question 2.4 On cherche à tester l'indépendance des variables croisées dans la **Question 1.4** (sexe et retard scolaire au moment de l'enquête). Le tableau de résultat est le suivant :

Statistic	DF	Value	Prob
Chi-Square	1	10.1644	0.0014
Likelihood Ratio Chi-Square	1	10.1642	0.0014
Continuity Adj. Chi-Square	1	9.7907	0.0018
Mantel-Haenszel Chi-Square	1	10.1570	0.0014
Phi Coefficient		0.0862	
Contingency Coefficient		0.0859	
Cramer's V		0.0862	

a. Rappelez comment est posé le test d'indépendance de deux variables qualitatives ainsi que le comportement de sa statistique de test D^2 sous l'hypothèse nulle.

b. Quelle est la valeur de cette statistique ? Vérifiez qu'il est bien possible de la retrouver à partir du seul tableau de la **Question 1.4**. Quelle option utiliser pour faire apparaître le tableau de résultat ci-dessus ?

c. En utilisant les quantiles de la loi du χ^2 en annexe, déterminez la valeur critique du test au seuil de 10 %. Que concluez-vous ?

d. Interprétez la p-valeur du test : est-il possible de rejeter l'hypothèse nulle à un seuil plus prudent que 10 % ? En aurait-il été de même si cette p-valeur avait valu 0,023 ?

Question 2.5 On cherche à tester la significativité de la corrélation entre les scores en mathématiques et en compréhension de l'écrit. L'ensemble des informations nécessaires figurent dans le tableau qui accompagne la **Question 1.5**.

a. Rappelez comment est posé le test d'indépendance de deux variables quantitatives ainsi que le comportement de sa statistique de test t sous l'hypothèse nulle.

b. Calculez la statistique de test et, en utilisant les quantiles de la loi de Student en annexe, menez le test correspondant au seuil de 1 %.

c. Par ailleurs, interprétez la p-valeur et concluez.

Session 3 : Analyse de variance

Question 3.1 Hypothèses de l'ANOVA Dans cette question, on souhaite tester sur les données de l'enquête PISA 2012 les hypothèses de normalité et d'homogénéité. Les résultats des tests de Shapiro-Wilk et de Bartlett menés sur l'ANOVA du score synthétique en mathématiques (PV1MATH) selon le sexe (ST04Q01) sont les suivants :

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.99814	**Pr < W**	0.1322
Kolmogorov-Smirnov	**D**	0.020463	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.074578	**Pr > W-Sq**	0.2464
Anderson-Darling	**A-Sq**	0.474044	**Pr > A-Sq**	0.2453

Bartlett's Test for Homogeneity of PV1MATH Variance			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
ST04Q01	1	4.6070	0.0318

a. Rappelez l'hypothèse nulle et l'hypothèse alternative de ces deux tests.

b. Commentez la p-valeur du test de Shapiro-Wilk : que concluez-vous ? Êtes-vous surpris du résultat (en repensant aux questions de la session 1) ?

c. Commentez la statistique de test et la p-valeur du test de Bartlett. Que concluez-vous ? Quelle est la conséquence de ce résultat sur l'analyse de la variance ?

Question 3.2 ANOVA selon un facteur dichotomique : TTEST On cherche à tester l'égalité des moyennes du score synthétique en mathématiques (PV1MATH) selon le sexe (ST04Q01). Les résultats sont les suivants :

Method	**Variances**	**DF**	**t Value**	**Pr > |t|**
Pooled	Equal	1366	-1.56	0.1181
Satterthwaite	Unequal	1327.9	-1.56	0.1194

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	653	713	1.18	0.0318

a. Quelle procédure a été utilisée pour produire ces résultats ?

b. Interprétez le test d'égalité des variances. Comparez avec les résultats de la **Question 3.1**.

c. Interprétez le test d'égalité des moyennes sous l'hypothèse d'inégalité des variances en utilisant la statistique de test. Vérifiez que votre conclusion est bien cohérente avec l'interprétation de la p-valeur.

d. Comparez avec le test d'égalité des moyennes sous l'hypothèse d'égalité des variances. Que retenez-vous d'un point de vue qualitatif ?

Question 3.3 ANOVA selon un facteur polytomique On cherche à analyser la variance du score synthétique en mathématiques (`PV1MATH`) selon le plus haut diplôme des parents (`hdipl`). Les résultats sont les suivants :

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1060169.54	353389.85	43.56	<.0001
Error	1289	10457452.19	8112.84		
Corrected Total	1292	11517621.72			

R-Square	Coeff Var	Root MSE	PV1MATH Mean
0.092048	17.89535	90.07131	503.3223

Source	DF	Anova SS	Mean Square	F Value	Pr > F
hdipl	3	1060169.538	353389.846	43.56	<.0001

a. Quelle procédure a été utilisée pour produire ces résultats ?

b. Rappelez la formule de la statistique de test utilisée dans le test de l'ANOVA. Repérez-la dans les sorties puis utilisez les autres résultats pour la recalculer.

c. On souhaite mener le test au seuil de 5 %. Quels sont les degrés de liberté de la loi de Fisher que suit la statistique de test sous H_0 ? Quelle est la valeur critique du test ? Que concluez-vous ? Vérifiez que votre conclusion est bien cohérente avec l'interprétation de la p-valeur.

Remarque : Si on procède à l'analyse de la variance du score synthétique en mathématiques selon le sexe en utilisant la même procédure que dans la **Question 3.3**, on obtient les résultats suivants (partiels) :

Source	DF	Anova SS	Mean Square	F Value	Pr > F
ST04Q01	1	22553.10330	22553.10330	2.45	0.1181

Comparez la p-valeur avec celles obtenues à la **Question 3.2**. Que remarquez-vous ? Comparez également les statistiques de test : remarquez-vous une relation entre t et F ?

Session 4 : Régression linéaire

L'objectif des questions de cette session est d'identifier certains déterminants des résultats au test standardisé de mathématiques (variable `PV1MATH`). Les variables explicatives sont intégrées une à une dans le modèle, dans l'ordre :

- le nombre d'heures de travail personnel consacré aux mathématiques chaque semaine : variable `mhours` ;
- le sexe : variable `ST04Q01` dichotomisée avec les variables `femme` et `homme` ;
- le plus haut niveau d'étude atteint par les parents : variable `hdipl` dichotomisée avec les variables `hdipl02`, `hdipl34`, `hdipl5` et `hdipl6` (*cf.* **Question 1.6** pour la signification des modalités de cette variable).

Question 4.1 Régression linéaire simple On estime tout d'abord le modèle :

$$\text{PV1MATH} = \beta_0 + \beta_1 \times \text{mhours} + u$$

dont les résultats sont les suivants :

Number of Observations Read	1368
Number of Observations Used	724
Number of Observations with Missing Values	644

Root MSE	89.90534	**R-Square**	0.0226
Dependent Mean	511.60180	**Adj R-Sq**	0.0212
Coeff Var	17.57330		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	476.46714	9.23439	51.60	<.0001
mhours	1	10.29129	2.52157	4.08	<.0001

a. Quelle procédure a été utilisée pour produire ces résultats ?

b. Que remarquez-vous quant au nombre d'observations effectivement utilisées dans le modèle ? Quelle est selon vous l'origine de ce phénomène ?

c. Quelle est la valeur du R^2 du modèle ? Est-ce une valeur faible ou une valeur élevée ?

d. Interprétez la valeur de $\hat{\beta}_1$. Sachant que la variance de `mhours` vaut 1,74 et la covariance de `PV1MATH` et `mhours` vaut 18,10, recalculez-la manuellement.

e. Rappelez l'hypothèse nulle et l'hypothèse alternative du test de significativité de la valeur du coefficient β_1. Quelle est sa statistique de test et quelle loi suit-elle sous l'hypothèse nulle ?

f. Menez le test de significativité de β_1 au seuil de 5 %. Que concluez-vous ? Vérifiez que cette conclusion est cohérente avec l'interprétation de la p-valeur du test.

Question 4.2 On intègre la variable de sexe au modèle :

$$\text{PV1MATH} = \beta_0 + \beta_1 \times \text{mhours} + \beta_2 \times \text{femme} + u$$

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	142478	71239	8.81	0.0002
Error	721	5828065	8083.30726		
Corrected Total	723	5970543			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	479.81434	9.84016	48.76	<.0001
mhours	1	10.32120	2.52181	4.09	<.0001
femme	1	-6.58919	6.69062	-0.98	0.3250

a. Rappelez l'hypothèse nulle et l'hypothèse alternative du test de significativité globale des coefficients d'une régression. Quelle est sa statistique de test et quelle loi suit-elle sous l'hypothèse nulle ?

b. Repérez la valeur de cette statistique de test et menez le test à un niveau de confiance de 95 %. Que concluez-vous ? Vérifiez que cette conclusion est bien cohérente avec l'interprétation de la p-valeur du test.

c. Pourquoi n'a-t-on pas intégré directement la variable de sexe (`ST04Q01`) dans le modèle ? Pourquoi la variable indicatrice `homme` n'apparaît-elle pas ?

d. Interprétez la valeur de $\hat{\beta}_2$. Ce coefficient est-il significativement différent de 0 aux seuils de 10 %, 5 %, 1 % ?

Question 4.3 On intègre au modèle la variable de plus haut niveau d'étude atteint par les parents par le biais de ses indicatrices :

$$\text{PV1MATH} = \beta_0 + \beta_1 \times \text{mhours} + \beta_2 \times \text{femme} + \beta_3 \times \text{hdipl02} + \beta_4 \times \text{hdipl5} + \beta_5 \times \text{hdipl6} + u$$

Root MSE	85.35541	**R-Square**	0.1239
Dependent Mean	511.60180	**Adj R-Sq**	0.1178
Coeff Var	16.68395		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	463.89103	10.08136	46.01	<.0001
mhours	1	9.38701	2.39872	3.91	<.0001
femme	1	-5.39314	6.37038	-0.85	0.3975
hdipl02	1	-45.01388	11.93042	-3.77	0.0002
hdipl5	1	25.66931	8.56069	3.00	0.0028
hdipl6	1	50.44090	7.67687	6.57	<.0001

a. Comparez le R^2 obtenu dans ce dernier modèle à celui du modèle linéaire simple de la **Question 4.1**. Comment expliquez-vous cette évolution ?

b. Pourquoi ne pas avoir intégré au modèle la variable indicatrice `hdipl34` ? Ce choix de modalité de référence vous paraît-il judicieux ?

c. Interprétez la valeur de $\hat{\beta}_3$, $\hat{\beta}_4$ et $\hat{\beta}_5$. Ces coefficients peuvent-ils être considérés comme statistiquement significatifs aux seuils statistiques usuels ?

Session 5 : Régression logistique

Dans cette session, on examine certaines variables susceptibles d'influencer le retard scolaire : sexe, diplôme des parents, conditions de vie. Les conditions de vie sont abordées à travers les variables `chambre`, `bureau`, `ordi` et `manuel` qui indiquent respectivement si la personne interrogée dispose d'une chambre individuelle, d'un bureau, d'un ordinateur et de manuels scolaires. On estime ainsi le modèle :

$$\begin{aligned} \text{retard} &= \beta_0 + \beta_1 \times \text{femme} + \beta_2 \times \text{hdipl02} + \beta_3 \times \text{hdipl5} + \beta_4 \times \text{hdipl6} \\ &\quad + \beta_5 \times \text{chambre} + \beta_6 \times \text{bureau} + \beta_7 \times \text{ordi} + \beta_8 \times \text{manuel} + u \end{aligned}$$

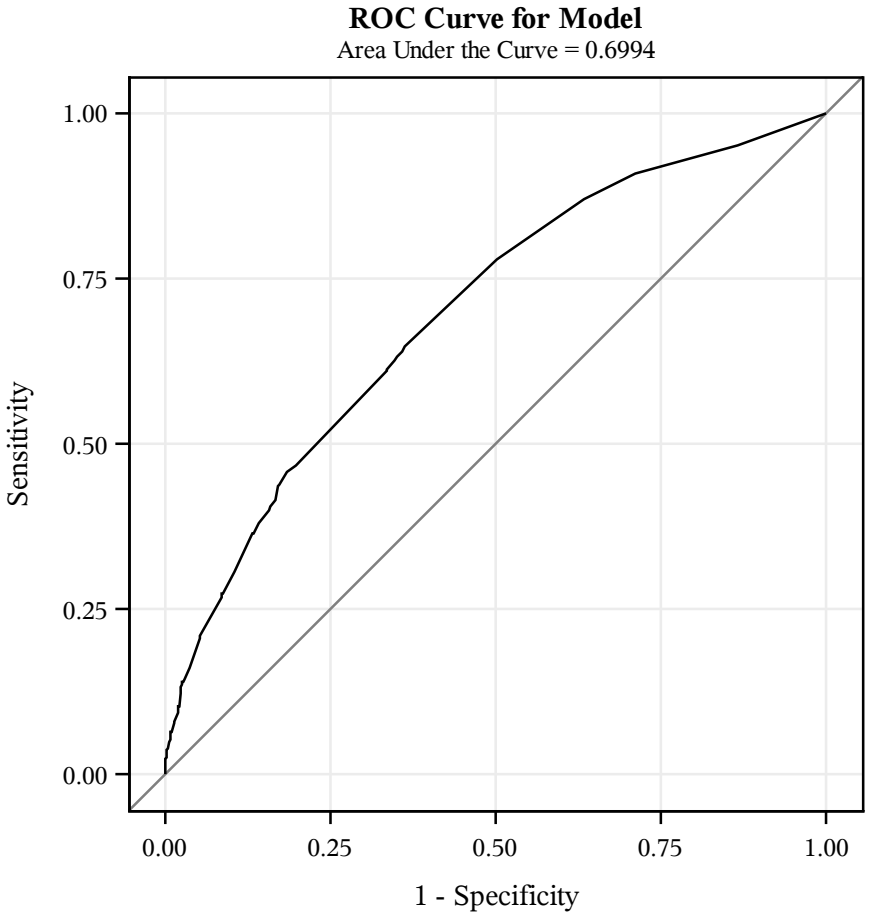
Question 5.1 Indicateurs de qualité du modèle Les indicateurs de qualité du modèle sont les suivants :

Number of Observations Read	1368
Number of Observations Used	1329

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	1601.784	1477.287
SC	1606.977	1524.017
-2 Log L	1599.784	1459.287

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	140.4972	8	<.0001
Score	144.1774	8	<.0001
Wald	121.5095	8	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	66.3	**Somers' D**	0.399
Percent Discordant	26.4	**Gamma**	0.430
Percent Tied	7.3	**Tau-a**	0.164
Pairs	363440	**c**	0.699



a. Comparez la valeur des différents indicateurs construits à partir de la log-vraisemblance. Êtes-vous en mesure de recalculez l'AIC et le SC à partir de -2 Log L ?

b. Identifiez la statistique du test de significativité globale par le ratio de vraisemblance. Êtes-vous en mesure de la recalculez à partir des autres informations de la sortie ? Interprétez ce test.

c. Quel est le pourcentage de concordance ? Vous paraît-il élevé ? Interprétez un des points de la courbe ROC. Que pensez-vous de son allure générale ?

Question 5.2 Interprétation des coefficients Ce modèle conduit à l'estimation des coefficients suivants :

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	2.8617	0.6259	20.9012	<.0001
femme	1	-0.3969	0.1305	9.2557	0.0023
hdipl02	1	0.8532	0.2099	16.5179	<.0001
hdipl5	1	-0.6175	0.1774	12.1178	0.0005
hdipl6	1	-0.5306	0.1606	10.9106	0.0010
chambre	1	-0.9312	0.1908	23.8313	<.0001
bureau	1	-0.5728	0.4541	1.5911	0.2072
ordi	1	-1.3739	0.4042	11.5540	0.0007
manuel	1	-0.7896	0.1777	19.7335	<.0001

a. Rappelez l'hypothèse nulle et l'hypothèse alternative du test de significativité des coefficients (par exemple pour β_1 associé à la variable `femme`). Quelle est la statistique de test et quelle loi suit-elle sous H_0 ? Menez le test au seuil de 5 %. Interprétez également la p-valeur.

b. Quelles sont les variables significatives aux seuils statistiques usuels ?

Question 5.3 Interprétation des *odds-ratio* Le modèle produit enfin le tableau d'*odds-ratio* suivant :

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
femme	0.672	0.521	0.868
hdipl02	2.347	1.555	3.542
hdipl5	0.539	0.381	0.764
hdipl6	0.588	0.429	0.806
chambre	0.394	0.271	0.573
bureau	0.564	0.232	1.373
ordi	0.253	0.115	0.559
manuel	0.454	0.320	0.643

a. Interprétez la valeur de l'*odds-ratio* associé à la variable `hdipl02` (référez-vous à la question 1.6 pour connaître la signification des modalités de la variable `hdipl`). Pouvez-vous déterminer si l'association entre faible diplôme des parents et retard scolaire est statistiquement significative à partir de ce tableau ?

b. Interprétez la valeur de l'*odds-ratio* associé au fait que l'individu dispose d'un bureau. L'association avec le retard scolaire est-elle significative ?

Session 6 : Compléments

Question 6.1 Introduire une variable et son carré À partir des données de l'EEC 2012T4 (données du support de révision), on estime le modèle :

$$\text{salaire} = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{age}^2 + \beta_3 \times \text{femme} + u$$

qui conduit à l'estimation (les erreurs-standards sont indiquées entre parenthèses) :

$$\text{salaire} = \underset{(422)}{-1473} + \underset{(21)}{160} \times \text{age} - \underset{(0,26)}{1,71} \times \text{age}^2 - \underset{(75)}{423} \times \text{femme} + u$$

Le modèle est estimé sur 647 observations et son R^2 est 0,1409.

a. Quel est le salaire prédit par le modèle pour une femme de 30 ans ? un homme de 50 ans ? Pouvez-vous déterminer l'âge pour lequel le salaire est maximal d'après le modèle ?

b. Quelle est la différence moyenne de salaire associée à une année supplémentaire à 30 ans ? à 50 ans ?

c. Posez le test de significativité jointe des coefficients β_1 et β_2. Quelle est la statistique traditionnellement associée à ce type de test et quelle loi suit-elle sous l'hypothèse nulle ?

d. Sachant que le R^2 du modèle contraint vaut 0,0334, quelle est la valeur de la statistique de test ? Que concluez-vous ?

e. Comment indiquer à SAS de mener directement ce test ?

Question 6.2 Introduire la variable explicative ou une variable expliquée en logarithme On estime plusieurs spécifications de modèles de régression linéaire simple faisant intervenir la variable expliquée salaire et la variable explicative age. Dans chaque cas, interprétez le coefficient de age.

a. $\text{salaire} = 984 + 19 \times \text{age} + u$

b. $ln(\text{salaire}) = 6,77 + 0,01 \times \text{age} + u$

c. $ln(\text{salaire}) = 5,05 + 0,61 \times ln(\text{age}) + u$

Annexe : Tables statistiques usuelles

Table 1 : Quantiles de la loi normale centrée réduite

Le quantile de niveau γ d'une variable aléatoire X suivant une loi normale centrée réduite noté $q_\gamma^{\mathcal{N}(0,1)}$ est défini par :

$$\Phi(q_\gamma^{\mathcal{N}(0,1)}) = \mathbb{P}(X \leq q_\gamma^{\mathcal{N}(0,1)}) = \gamma$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

γ	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**
$q_\gamma^{\mathcal{N}(0,1)}$	-2,58	-2,33	-1,96	-1,64	-1,28	1,28	1,64	1,96	2,33	2,58

Lecture : 97,5% des valeurs d'une variable aléatoire suivant une loi normale centrée réduite sont inférieures à 1,96.

Table 2 : Quantiles de la loi du χ^2

Le quantile de niveau γ d'une variable aléatoire X suivant une loi du χ^2 à p degrés de liberté noté $q_\gamma^{\chi^2_p}$ est défini par :

$$F_X(q_\gamma^{\chi^2_p}) = \mathbb{P}(X \leq q_\gamma^{\chi^2_p}) = \gamma$$

où F_X est la fonction de répartition de X.

p \ γ	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**
1	0,00	0,00	0,00	0,00	0,02	2,71	3,84	5,02	6,63	7,88
2	0,01	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	10,60
3	0,07	0,11	0,22	0,35	0,58	6,25	7,81	9,35	11,34	12,84
4	0,21	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55
7	0,99	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,95
9	1,73	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19
20	7,43	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00
30	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	37,69	63,17	67,50	71,42	76,15	79,49
100	67	70	74	78	82	118	124	130	136	140
1000	889	899	914	928	943	1058	1075	1090	1107	1119

Lecture : 95% des valeurs d'une variable aléatoire suivant une loi du χ^2 à 1 degré de liberté sont inférieures à 3,84.

Table 3 : Quantiles de la loi de Student

Le quantile de niveau γ d'une variable aléatoire X suivant une loi de Student à p degrés de liberté noté $F_X(q_\gamma^{\mathcal{T}_p}) = q_\gamma^{\mathcal{T}_p}$ est défini par :

$$F(q_\gamma^{\mathcal{T}_p}) = \mathbb{P}(X \leq q_\gamma^{\mathcal{T}_p}) = \gamma$$

où F_X est la fonction de répartition de X.

p \ γ	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**
1	-63,66	-31,82	-12,71	-6,31	-3,08	3,08	6,31	12,71	31,82	63,66
2	-9,92	-6,96	-4,30	-2,92	-1,89	1,89	2,92	4,30	6,96	9,92
3	-5,84	-4,54	-3,18	-2,35	-1,64	1,64	2,35	3,18	4,54	5,84
4	-4,60	-3,75	-2,78	-2,13	-1,53	1,53	2,13	2,78	3,75	4,60
5	-4,03	-3,36	-2,57	-2,02	-1,48	1,48	2,02	2,57	3,36	4,03
6	-3,71	-3,14	-2,45	-1,94	-1,44	1,44	1,94	2,45	3,14	3,71
7	-3,50	-3,00	-2,36	-1,89	-1,41	1,41	1,89	2,36	3,00	3,50
8	-3,36	-2,90	-2,31	-1,86	-1,40	1,40	1,86	2,31	2,90	3,36
9	-3,25	-2,82	-2,26	-1,83	-1,38	1,38	1,83	2,26	2,82	3,25
10	-3,17	-2,76	-2,23	-1,81	-1,37	1,37	1,81	2,23	2,76	3,17
20	-2,85	-2,53	-2,09	-1,72	-1,33	1,33	1,72	2,09	2,53	2,85
30	-2,75	-2,46	-2,04	-1,70	-1,31	1,31	1,70	2,04	2,46	2,75
40	-2,70	-2,42	-2,02	-1,68	-1,30	1,30	1,68	2,02	2,42	2,70
50	-2,68	-2,40	-2,01	-1,68	-1,30	1,30	1,68	2,01	2,40	2,68
100	-2,63	-2,36	-1,98	-1,66	-1,29	1,29	1,66	1,98	2,36	2,63
1000	-2,58	-2,33	-1,96	-1,65	-1,28	1,28	1,65	1,96	2,33	2,58
$+\infty$	-2,58	-2,33	-1,96	-1,64	-1,28	1,28	1,64	1,96	2,33	2,58

Lecture : 97,5% des valeurs d'une variable aléatoire suivant une loi de Student à 1 degré de liberté sont inférieures à 12,71. On remarque que quand p tend vers $+\infty$, les quantiles de la loi de Student sont ceux de la loi normale centrée réduite.

Table 4 : Quantiles de la loi de Fisher

Le quantile de niveau γ d'une variable aléatoire X suivant une loi de Fisher à q et p degrés de liberté noté $q_\gamma^{F_{q,p}}$ est défini par :

$$F_X(q_\gamma^{F_{q,p}}) = \mathbb{P}(X \leq q_\gamma^{F_{q,p}}) = \gamma$$

où F_X est la fonction de répartition de X.

On présente les quantiles à 0,95 et 0,99 d'une loi de Fisher pour les degrés de liberté q et p usuels.

Quantiles de niveau $\gamma = 0,95$ d'une loi de Fisher $F_{q,p}$ à q et p degrés de liberté.

p \ q	1	2	3	4	5	6	7	8	9	10
1	161	199	216	225	230	234	237	239	241	242
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93
1000	3,85	3,00	2,61	2,38	2,22	2,11	2,02	1,95	1,89	1,84

Lecture : 95% des valeurs d'une variable aléatoire suivant une loi de Fisher $F_{1,100}$ inférieures à 3,94.

Quantiles de niveau $\gamma = 0,99$ d'une loi de Fisher $F_{q,p}$ à q et p degrés de liberté.

p \ q	1	2	3	4	5	6	7	8	9	10
1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50
1000	6,66	4,63	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34

Lecture : 99% des valeurs d'une variable aléatoire suivant une loi de Fisher $F_{1,100}$ sont inférieures à 6,90.