Section 02 : Régression et Moindres Carrés Ordinaires (Application Stata)

GSF-6053: Économétrie Financière

Simon-Pierre Boucher[1]

[1]Département de finance, assurance et immobilier
Faculté des sciences de l'administration
Université Laval

1 Février 2022

Application STATA

- Dans cette application Stata, nous allons analyser le jeu de données portant le nom **Wage1**
- Il s'agit essentiellement de données collectées auprès de plusieurs travailleurs en 1976.
- Dans cette analyse nous chercherons à quantifier et expliquer l'impact de certains facteurs propre au salarié sur son salaire.
- Les variables que nous utiliserons dans cette analyse sont les suivantes:
  - **lwage:** log du salaire horaire moyen
  - **educ:** années d'études
  - **exper:** années d'expérience potentielle
  - **tenutre:** années chez l'employeur actuel

Application STATA

Code 1: Pour obtenir jeu de données Wage1 sur Stata

use http://fmwww.bc.edu/ec-p/data/wooldridge/wage1

Code 2: Vérifier le format des variables utilisées

describe wage educ exper tenure

Output 2:

variable name	storage type	display format
wage	float	%9.0g
educ	float	%9.0g
exper	float	%9.0g
tenure	float	%9.0g

- On peut voir que nos trois variables sont de type Float, ce qui indique qu'il s'agit d'une valeur numérique entière.
- On peut donc l'utiliser pour notre analyse.

Application STATA

Code 3: Calculer et afficher des statistiques descriptives

summarize wage educ exper tenurex

Output 3:

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	526	5.896103	3.693086	.53	24.98
educ	526	12.56274	2.769022	0	18
exper	526	17.01711	13.57216	1	51
tenure	526	5.104563	7.224462	0	44

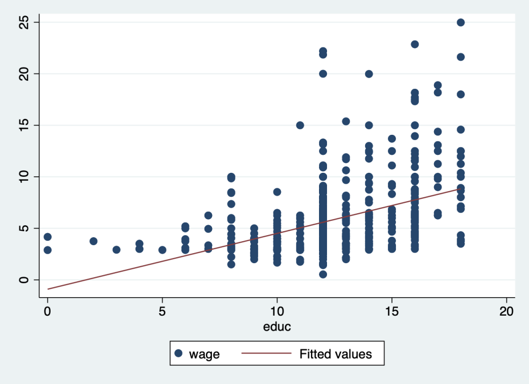
- On peut voir que notre jeu de données contient 526 observations
- Que le salaire moyen horraire est de 5.89 et qu'en moyenne les salariés ont 12.56 années de scolarité.
- Pour ce qui est de l'expérience de travail, en moyenne les salariés ont 17.02 années d'expériences et ils sont dans le même emploie depuis 5.10 ans, en moyenne.

Application STATA

Code 4: Relation entre le salaire et l'éducation

twoway (scatter wage educ) (lfit wage educ)

Output 4:

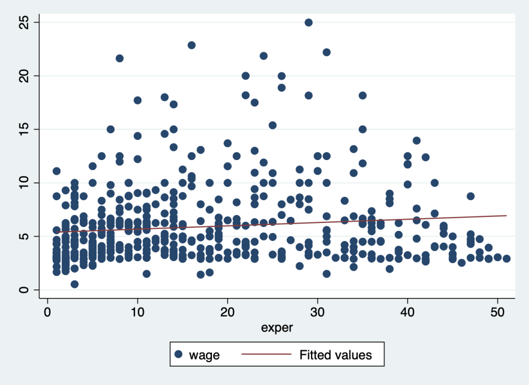


Application STATA

Code 5: Relation entre le salaire et l'expérience de travail

twoway (scatter wage exper) (lfit wage exper)

Output 5:

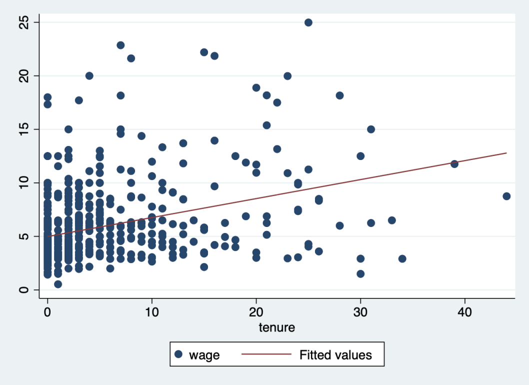


Application STATA

Code 6: Relation entre le salaire et le nombre d'années chez l'employeur actuel

twoway (scatter wage tenure) (lfit wage tenure)

Output 6:

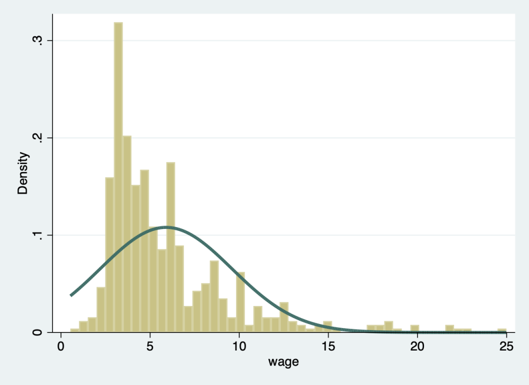


Application STATA

Code 7: Distrbution de la variable **wage**

histogram wage, bin(50) normal normopts(lwidth(thick))

Output 7:

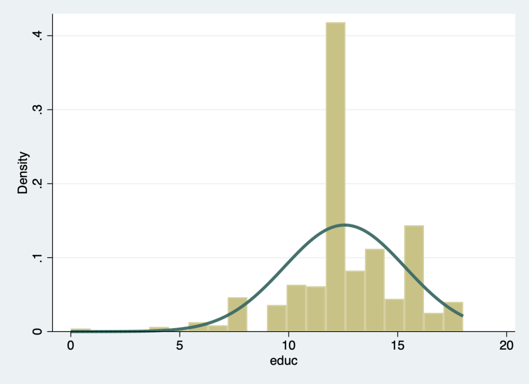


Application STATA

Code 8: Distrbution de la variable **educ**

```
histogram educ, bin(20) normal normopts(lwidth(thick))
```

Output 8:

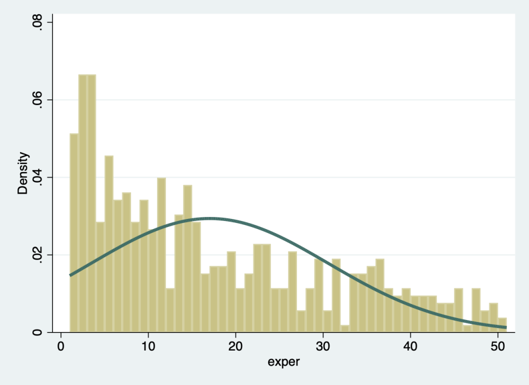


Application STATA

Code 9: Distrbution de la variable **exper**

histogram exper, bin(50) normal normopts(lwidth(thick))

Output 9:

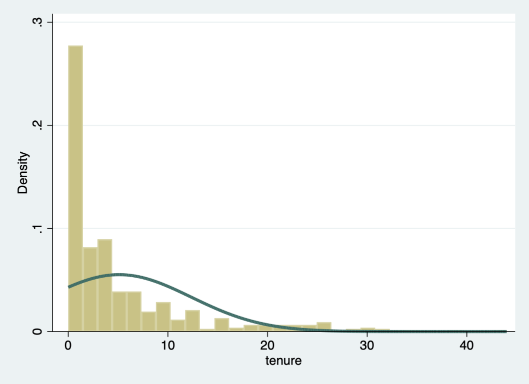


Application STATA

Code 10: Distrbution de la variable **tenure**

histogram tenure, bin(30) normal normopts(lwidth(thick))

Output 10:



Application STATA

Régression linéaire simple

$$wage = \beta_0 + \beta_1 educ + \epsilon \tag{1}$$

$$wage = \beta_0 + \beta_1 exper + \epsilon \tag{2}$$

$$wage = \beta_0 + \beta_1 tenure + \epsilon \tag{3}$$

Régression linéaire multiple

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon \tag{4}$$

Application STATA

Code 11: $wage = \beta_0 + \beta_1 educ + \epsilon$

reg wage educ

Output 11:

Source	SS	df	MS
Model	1179.73204	1	1179.73204
Residual	5980.68225	524	11.4135158
Total	7160.41429	525	13.6388844

Number of obs	=	526
F(1, 524)	=	103.36
Prob > F	=	0.0000
R-squared	=	0.1648
Adj R-squared	=	0.1632
Root MSE	=	3.3784

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.5413593	.053248	10.17	0.000	.4367534	.6459651
_cons	-.9048516	.6849678	-1.32	0.187	-2.250472	.4407687

Application STATA

Analyse: $wage = \beta_0 + \beta_1 educ + \epsilon$

- Avec un R^2 de 0.1648, le nombre d'années d'éducation d'un salarié semble expliquer environ 16% des variations dans le salaire de ce dernier, et ce en moyenne.
- La p-value du T-test pour le coefficient de la variable educ est de 0.000, ce qui implique que le coefficient est significativement différent de zéro. (À un seuil de 5%)
- Avec un coefficient estimé de 0.54 pour la variable educ, on peut comprendre qu'en moyenne, une année d'éducation additionnelle amène une augmentation du salaire horaire de 0.54.

Application STATA

Code 12: $wage = \beta_0 + \beta_1 exper + \epsilon$

reg wage exper

Output 12:

Source	SS	df	MS
Model	91.2751351	1	91.2751351
Residual	7069.13916	524	13.4907236
Total	7160.41429	525	13.6388844

Number of obs	=	526
F(1, 524)	=	6.77
Prob > F	=	0.0096
R-squared	=	0.0127
Adj R-squared	=	0.0109
Root MSE	=	3.673

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	.0307219	.0118111	2.60	0.010	.007519	.0539247
_cons	5.373305	.2569919	20.91	0.000	4.868444	5.878166

Application STATA

Analyse: $wage = \beta_0 + \beta_1 exper + \epsilon$

- Avec un R^2 de 0.0127, le nombre d'années d'expérience d'un salarié semble expliquer environ 1% des variations dans le salaire de ce dernier, et ce en moyenne.
- La p-value du T-test pour le coefficient de la variable exper est de 0.01, ce qui implique que le coefficient est significativement différent de zéro. (À un seuil de 5%)
- Avec un coefficient estimé de 0.0307 pour la variable exper, on peut comprendre qu'en moyenne, une année d'expérience additionnelle amène une augmentation du salaire horaire de 0.0307.

Application STATA

Code 13: $wage = \beta_0 + \beta_1 tenure + \epsilon$

reg wage tenure

Output 13:

Source	SS	df	MS			
				Number of obs	=	526
				F(1, 524)	=	71.68
Model	861.62965	1	861.62965	Prob > F	=	0.0000
Residual	6298.78464	524	12.0205814	R-squared	=	0.1203
				Adj R-squared	=	0.1187
Total	7160.41429	525	13.6388844	Root MSE	=	3.4671

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tenure	.1773271	.0209449	8.47	0.000	.1361809	.2184733
_cons	4.990925	.185158	26.95	0.000	4.627182	5.354669

Application STATA

Analyse: $wage = \beta_0 + \beta_1 tenure + \epsilon$

- Avec un R^2 de 0.1203, le nombre d'années chez l'employeur actuel d'un salarié semble expliquer environ 12% des variations dans le salaire de ce dernier, et ce en moyenne.
- La p-value du T-test pour le coefficient de la variable tenure est de 0.000, ce qui implique que le coefficient est significativement différent de zéro. (À un seuil de 5%)
- Avec un coefficient estimé de 0.177 pour la variable tenure, on peut comprendre qu'en moyenne, une année de plus chez l'employeur actuel amène une augmentation du salaire horaire de 0.177.

Application STATA

Code 14: $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$

reg wage educ exper tenure

Output 14:

Source	SS	df	MS
Model	2194.1116	3	731.370532
Residual	4966.30269	522	9.51398984
Total	7160.41429	525	13.6388844

Number of obs	=	526
F(3, 522)	=	76.87
Prob > F	=	0.0000
R-squared	=	0.3064
Adj R-squared	=	0.3024
Root MSE	=	3.0845

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.5989651	.0512835	11.68	0.000	.4982176	.6997126
exper	.0223395	.0120568	1.85	0.064	-.0013464	.0460254
tenure	.1692687	.0216446	7.82	0.000	.1267474	.2117899
_cons	-2.872735	.7289643	-3.94	0.000	-4.304799	-1.440671

Application STATA

Analyse: $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$

- Avec un R^2 de 0.3064, les trois variables combinées ensemble semblent expliquer environ 31% des variations dans le salaire de ce dernier, et ce en moyenne.
- De plus, avec un R^2 ajusté de 0.3024 (relativement proche du R^2 standard), l'ajout de variables explicatives semble justifier étant donné que la valeur du R^2 ajusté ne semble pas avoir diminuer grandement par rapport au R^2 standard, suite à la pénalité pour l'ajout de variables explicatives dans le R^2 ajusté.
- La p-value du F-test est de 0.000, ce qui implique que nous pouvons rejeter l'hypothèse nulle que tous les coefficients sont conjointement non significatifs.

Détails sur le T-test

- Dans le T-test que nous venons de faire, nous avons utilisé la P-value pour vérifier si le coefficient est significativement différent de 0.
- En effet si cette P-value est inférieur à 0.05, on rejette l'hypothèse nulle que le coefficient est pas significativement différent de 0.
- Afin de bien comprendre comment le T-test fonctionne, il est important de bien faire l'analyse sans utiliser la P-value.
- Nous allons reprendre la dernière régression que nous venons d'analyser, soit celle avec toutes les variables explicatives incluses.

Détails sur le T-test

Test de Student:

Two-tailed test

- Hypothèse nulle est la non-significativité du coefficient de régression
- Hypothèses:
  - $H_0 : \beta_k = 0 \rightarrow$ **Hypothèse nule**
  - $H_1 : \beta_k \neq 0 \rightarrow$ **Hypothèse alternative**
- Règle de décision: Rejeter H_0 si:

$$t = \frac{\mid \hat{\beta}_k - \beta_0 \mid}{SE_{\hat{\beta}_k}} > t_{n-k,\alpha/2}$$

  - Ou β_0 est la valeur du coefficient sous l'hypothèse nulle, soit 0
  - $SE_{\hat{\beta}_k}$ est l'écart-type associé à l'estimation de $\hat{\beta}_k$

Détails sur le T-test

- Le Modèle de régression:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$$

- Afin de pouvoir faire la comparaison de notre statistique T, nous devons obtenir une valeur critique à un seuil de 5%, soit

$$t_{n-k,\alpha/2}$$

- Sachant que nous avons $n = 526$ observations et $k = 3$ variables explicatives, nous devrons trouve la valeur critique:

$$t_{526-3,0.05/2} = t_{523,0.025}$$

Détails sur le T-test

- Nous allons maintenant aller regarder dans une table statistique appelée **T-table** afin de trouver notre valeur critique.
- Au niveau de la significativité, on prendra la colonne avec une valeur:
  - $\alpha = 0.025$ pour le **One-tail test**
  - $\alpha = 0.05$ pour le **Two-tail test**
- Pour ce qui est du nombre de degrés de liberté, nous avons $df = 523$, cependant la table ne fournit pas nécessairement une valeur pour tous les df possibles.
  - Dans la table qui va suivre, nous avons seulement $df = 100$ et $df = 1000$ de dispo, nous allons donc prendre $df = 1000$ par sécurité.

Détails sur le T-test

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
					Confidence Level						

Détails sur le T-test

- Notre valeur critique est de $t_{526-3,0.05/2} = t_{523,0.025} = 1.962$
- Pour le coefficient β_1 attaché à la variable **educ**:

$$t = \frac{\mid 0.5989651 - 0 \mid}{0.0512835} = 11.69 > 1.962$$

  - $t = 11.69$ est plus grand que la valeur critique $t_{523,0.025} = 1.962$, alors on peut rejetter l'hypothèse nulle et conclure que $\hat{\beta}_1$ est significativement différent de 0, à un seuil de 5%.

Détails sur le T-test

- Pour le coefficient β_2 attaché à la variable **exper**:

$$t = \frac{\mid 0.0223395 - 0 \mid}{0.0120568} = 1.85 < 1.962$$

  - $t = 1.85$ est plus petit que la valeur critique $t_{523,0.025} = 1.962$, alors on ne peut pas rejetter l'hypothèse nulle. $\hat{\beta}_2$ n'est pas significativement différent de 0, à un seuil de 5%.

Détails sur le T-test

- Pour le coefficient β_3 attaché à la variable **tenure**:

$$t = \frac{\mid 0.1692687 - 0 \mid}{0.0216446} = 7.82 > 1.962$$

  - $t = 7.82$ est plus grand que la valeur critique $t_{523,0.025} = 1.962$, alors on peut rejetter l'hypothèse nulle et conclure que $\hat{\beta}_3$ est significativement différent de 0, à un seuil de 5%.

Détails sur le F-test

- Comme pour le T-test, on veut également effectuer le F-test sans la P-value.

F-test (significativité conjointe)

Hypothèses:

- $H_O : \beta_2 = 0$ et/ou $\beta_3 = 0$ et/ou ... et/ou $\beta_k = 0$
- $H_1 : \beta_2 \neq 0$ et $\beta_3 \neq 0$ et ... et/ou $\beta_k \neq 0$

Statistique de test:

$$F = \frac{MS_{group}}{MS_{error}} > F(q, t-k; \alpha)$$

Sachant

- $MS_{group} = \frac{SS_{group}}{df_{group}}$
- $MS_{error} = \frac{SS_{error}}{df_{error}}$

Détails sur le F-test

- On rejette H_0 si la statistique F est supérieur à la valeur critique $F(q, t-k; \alpha)$.
- Dans le cas qui nous concerne, nous avons 4 coefficients (k=4), soit $\beta_0, \beta_1, \beta_2$ et β_3
- $q = k - 1 = 4 - 1 = 3 = df_{group}$
- $t - k = 526 - 4 = 522 = df_{error}$
- On pose un seuil de significativité de $\alpha = 0.05$
- La valeur critique sera représenté par $F(3, 522; 0.05)$
- Nous allons maintenant regarder dans la table suivante la valeur critique $F(3, 522; 0.05)$

Détails sur le F-test

Table A4: 5% Critical Values of the *F* Distribution

		Numerator Degrees of Freedom									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Denominator Degrees of Freedom	10	4.96	4.1	[illegible]	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	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
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	120	3.92	3.07	[illegible]	2.45	2.29	2.17	2.09	2.02	1.96	1.91
	∞	3.84	3.0	$\boxed{2.60}$	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Détails sur le F-test

- La valeur de la statistique F est:

$$F(3, 522; 0.05) = 2.60$$

- Voici maintenant le test:

$$F = \frac{\frac{SS_{group}}{df_{group}}}{\frac{SS_{error}}{df_{error}}} > F(q, t-k; \alpha)$$

$$F = \frac{\frac{2194.1116}{3}}{\frac{4966.30269}{522}} = 76.87 > 2.60$$

Détails sur le F-test

- On rejette l'hypothèse nulle étant donné que la statistique $F = 76.87$ est supérieur à la valeur critique $F(3, 522; 0.05) = 2.60$.
- On rejette donc l'hypothèse que tous les coefficients sont conjointement tous égaux à 0.
- En d'autres mots, au moins un coefficient dans cette régression est significativement différent de 0.

Application STATA

Code 15: Générer les résidus de la régression

```
reg wage educ exper tenure
predict resid, residuals
```

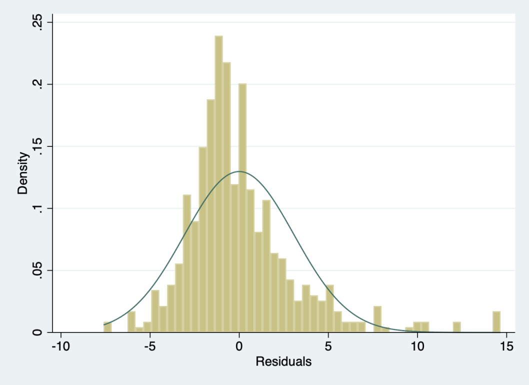
- Sans faire une analyse détaillée des résidus, nous allons créer une variable représentant les résidus de notre régression.
- Afin de faire l'analyse, nous allons produire un histogramme de nos résidus et un nuage de points avec les résidus en axe des Y et la variable **wage** en axe des X.

Application STATA

Code 16: Histogramme des résidus

histogram resid, bin(50) normal

Output 16:

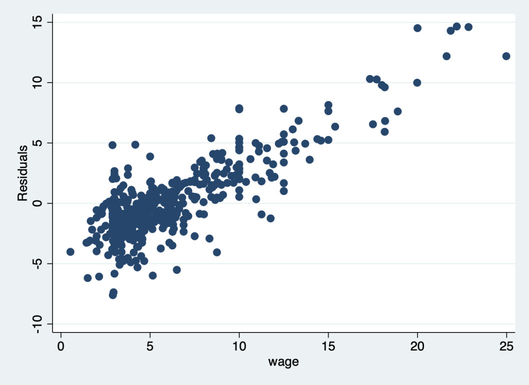


Application STATA

Code 17: Nuage de points des résidus

twoway (scatter red wage)

Output 17:



Application STATA

Analyse des résidus

- Comme caractéristiques souhaitez pour nos résidus, nous souhaitons qu'il ressemble à une loi normale et qu'il n'ait pas de **patterns** entre ce résidu et la variable **wage**
- Au niveau de l'histogramme des résidus, on peut voir que sans être parfaits les résidus prennent la forme d'une cloche.
- Il semble y avoir plus de résidus au niveau du centre (x=0), que ce qui est prévu par la loi normale.
- Au niveau du nuage de point, nous avons un problème majeur, il y a clairement une relation entre notre résidu et la variable **wage**.