Applied Statistics and Econometrics
Lecture 6

Giuseppe Ragusa

Luiss University

gragusa@luiss.it
http://gragusa.org/

March 6, 2017

Empirical application.

Data

Italian Labour Force Survey, ISTAT (2015Q3)

- wage: wage of full-time workers
- education: years of education

Wage and Education

Italia Labour Force Survey - ISTAT, 2015
3 Variables 26127 Observations

RETRIC

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
26127	0	275	0.999	1307	567.5	500	680	1000	1300	1550	1950	2290

```
lowest :  250  260  270  280  290, highest: 2960 2970 2980 2990 3000
```

EDULEV

n	missing	distinct
26127	0	6

Value	No education	elementary school	middle school
Frequency	142	700	7510
Proportion	0.005	0.027	0.287

Value	prof. high school diploma	high school diploma	college degree
Frequency	2289	10530	4956
Proportion	0.088	0.403	0.190

SG11

n	missing	distinct	Info	Mean	Gmd
26127	0	2	0.748	1.473	0.4985

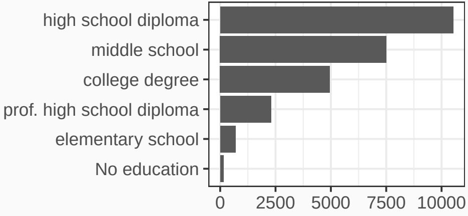
Wage and education: data

We recode education in terms of year of education

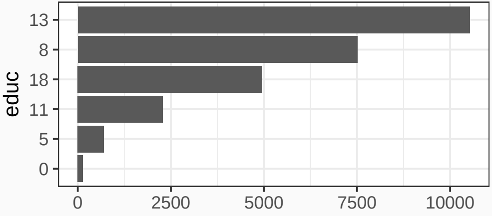
```
years <- c(0, 5, 8, 11, 13, 18)
cnt <- 1
for (j in levels(lfs$EDULEV)) {
    lfs$educ[lfs$EDULEV == j] <- years[cnt]
    cnt <- cnt + 1
}

table(lfs$educ)


##
##     0     5     8    11    13    18
##   142   700  7510  2289 10530  4956
```

Wage and education: data

Education levels



Years of education



Regression: wages and education

```
lm1 <- lm(RETRIC ~ educ, data = lfs)
summary(lm1)
```

```
##
## Call:
## lm(formula = RETRIC ~ educ, data = lfs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1312.6  -312.6   -32.5   252.4  1867.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  788.421     10.368    76.0   <2e-16 ***
## educ          43.012      0.822    52.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 497 on 26125 degrees of freedom
## Multiple R-squared:  0.0949,Adjusted R-squared:  0.0949
```

Regression when X is Binary (Section 5.3)

- $X = 1$ if small class size, $= 0$ if not;
- $X = 1$ if female, $= 0$ if not;
- etc.
- Binary regressors are sometimes called dummy variables.
- So far, β_1 has been called a "slope", but that doesnt make sense if X is binary.
- How do we interpret regression with a binary regressor?

Interpretation when X is binary

Consider

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

when X is binary.

Interpretation when X is binary

Consider

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

when X is binary. Then,

$$E[Y_i|X_i = 0] = \beta_0$$

$$E[Y_i|X_i = 1] = \beta_0 + \beta_1$$

Interpretation when X is binary

Consider

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

when X is binary. Then,

$$E[Y_i|X_i = 0] = \beta_0$$

$$E[Y_i|X_i = 1] = \beta_0 + \beta_1$$

Thus,

$$\begin{aligned} \beta_1 &= E[Y_i|X_i = 1] - E[Y_i|X_i = 0] \\ &= \text{population difference in group means} \end{aligned}$$

Example

Let

$$D_i = \begin{cases} 1 & \text{if } STR_i \leqslant 20 \\ 0 & \text{if } STR_i > 20 \end{cases}$$

```r
library(ase)
data(CASchools)
```

The linear model:

$$TestScore_i = \beta_0 + \beta_1 D_i + u_i$$

```r
CASchools["D"] <- ifelse(CASchools[["str"]] <= 20, 1, 0)
## OLS
lm(testscore ~ D, data = CASchools)

##
## Call:
## lm(formula = testscore ~ D, data = CASchools)
##
## Coefficients:
## (Intercept)            D
##      650.00         7.19
```

Difference in means/regression

	testscore		
D	n	mean	sd
0	177	649.999	17.966
1	243	657.185	19.286
All	420	654.157	19.053

$$\bar{Y}_{small} - \bar{Y}_{large} = 657.185 - 649.999$$
$$= 7.186 \approx 7.19$$

$$SE(\bar{Y}_{small} - \bar{Y}_{large}) = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}$$
$$= 1.83$$

```
summary(lm(testscore ~ D, data = CASchools))


Call:
lm(formula = testscore ~ D, data = CASchools)

Residuals:
   Min      1Q  Median      3Q     Max
-50.43 -14.07  -0.28  12.78  49.57

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   650.00       1.41  461.41  < 2e-16 ***
D               7.19       1.85    3.88  0.00012 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19 on 418 degrees of freedom
Multiple R-squared:  0.0348,Adjusted R-squared:  0.0324
F-statistic: 15.1 on 1 and 418 DF,  p-value: 0.000121
```

Difference in wages: males / females

```r
## SG11 denote gender of individual
## SG11 is coded as:
## == 1, male;
## == 2, female
## `female` is == 1 if female; ==0 o/w
lfs$female <- ifelse(lfs$SG11 == 2, 1, 0)
```

```r
lm(RETRIC ~ female, data = lfs)
```

```
##
## Call:
## lm(formula = RETRIC ~ female, data = lfs)
##
## Coefficients:
## (Intercept)       female
##        1445         -291
```

Heteroskedasticity and Homoskedasticity

Heteroskedasticity robust standard errors (Section 5.4)

- What...?
- Consequences of heteroskedasticicty/homoskedasticity
- Implication for computing standard errors

What do these two terms mean?

If $var(u|X = x)$ is constant — that is, if the variance of the conditional distribution of u given X does not depend on X then u is said to be homoskedastic. Otherwise, u is heteroskedastic.

Consider

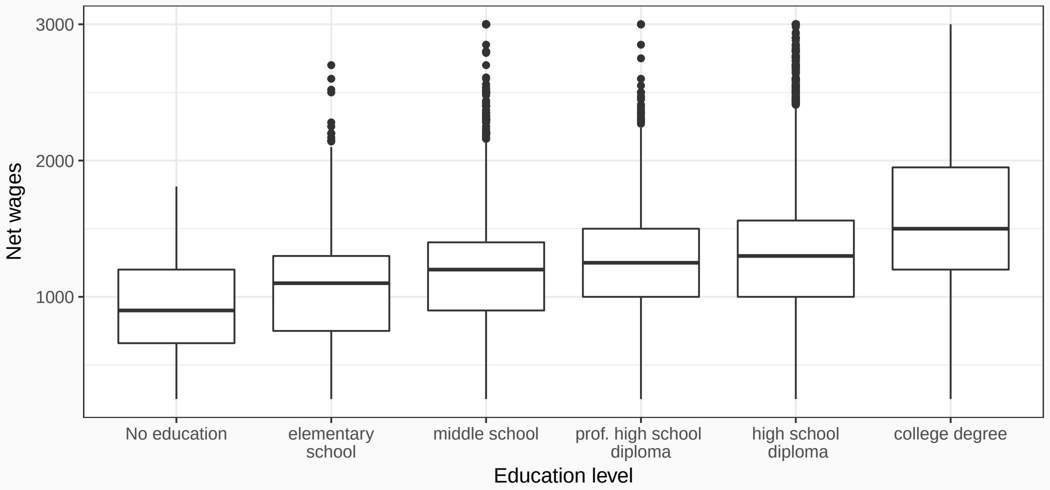
$$wage_i = \beta_0 + \beta_1 educ_i + u_i$$

Homoskedasticity means that the variance of u_i does not change with the education level.

Of course, we do not know anything about $var(u_i|educ_i)$, but we can use data to get an idea.

We plot the boxplot of wage for each educational cathegory — if u_i is homeskedastic, the box should approximately be of the same size

Homoskedasticity in a picture



Homoskedasticity in a table

Table 1: Sample variance of wage per education level

EDULEV	VAR	SD
No education	104210	323
elementary school	179437	424
middle school	184808	430
prof. high school diploma	184084	429
high school diploma	235692	485
college degree	401217	633

So far we have (without saying so) assumed that u might be heteroskedastic.

Recall the three least squares assumptions:

- $E(u|X = x) = 0$;
- (Xi, Yi), $i = 1, \ldots, n$, are $i.i.d.$
- Large outliers are rare

Heteroskedasticity and homoskedasticity concern $var(u|X = x)$. Because we have not explicitly assumed homoskedastic errors, we have implicitly allowed for heteroskedasticity.

Standard Errors

We now have two formulas for standard errors for $\hat{\beta}_1$:

- **Homoskedastic only standard errors**—these are valid only if the errors are homoskedastic
- The **heteroskedasticity robust standard errors** valid whether or not the errors are heteroskedastic.
- The main advantage of the homoskedasticity-only standard errors is that the formula is simpler. But the disadvantage is that the formula is only correct if the errors are homoskedastic.

Practical implications

- The homoskedasticity-only formula for the standard error of $\hat{\beta}_1$ and the heteroskedasticity-robust formula differ - so in general, you get different standard errors using the different formulas.

Practical implications

- The homoskedasticity-only formula for the standard error of $\hat{\beta}_1$ and the heteroskedasticity-robust formula differ - so in general, you get different standard errors using the different formulas.
- Homoskedasticity-only standard errors are the default setting in regression software - sometimes the only setting (e.g. Excel). To get the general heteroskedasticity-robust standard errors you must override the default.

Practical implications

- The homoskedasticity-only formula for the standard error of $\hat{\beta}_1$ and the heteroskedasticity-robust formula differ - so in general, you get different standard errors using the different formulas.
- Homoskedasticity-only standard errors are the default setting in regression software - sometimes the only setting (e.g. Excel). To get the general heteroskedasticity-robust standard errors you must override the default.
- If you dont override the default and there is in fact heteroskedasticity, your standard errors (and wrong t-statistics and confidence intervals) will be wrong - typically, homoskedasticity-only SEs are too small.

The bottom line...

- If the errors are either homoskedastic or heteroskedastic and you use heteroskedastic-robust standard errors, you are OK

The bottom line...

- If the errors are either homoskedastic or heteroskedastic and you use heteroskedastic-robust standard errors, you are OK
- If the errors are heteroskedastic and you use the homoskedasticity-only formula for standard errors, your standard errors will be wrong (the homoskedasticity-only estimator of the variance of $\hat{\beta}_1$ is inconsistent if there is heteroskedasticity).

The bottom line...

- If the errors are either homoskedastic or heteroskedastic and you use heteroskedastic-robust standard errors, you are OK
- If the errors are heteroskedastic and you use the homoskedasticity-only formula for standard errors, your standard errors will be wrong (the homoskedasticity-only estimator of the variance of $\hat{\beta}_1$ is inconsistent if there is heteroskedasticity).
- The two formulas coincide (when n is large) in the special case of homoskedasticity

The bottom line...

- If the errors are either homoskedastic or heteroskedastic and you use heteroskedastic-robust standard errors, you are OK
- If the errors are heteroskedastic and you use the homoskedasticity-only formula for standard errors, your standard errors will be wrong (the homoskedasticity-only estimator of the variance of $\hat{\beta}_1$ is inconsistent if there is heteroskedasticity).
- The two formulas coincide (when n is large) in the special case of homoskedasticity
- So, you should always use heteroskedasticity-robust standard errors

In `R`, to obtain heteroskedastic robust standard errors use

`summary_rob()`

```r
summary(lm(testscore ~ str, data = CASchools))
```

```

Call:
lm(formula = testscore ~ str, data = CASchools)

Residuals:
   Min     1Q Median     3Q    Max
-47.73 -14.25   0.48  12.82  48.54

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   698.93       9.47   73.82  < 2e-16
str            -2.28       0.48   -4.75  2.8e-06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19 on 418 degrees of freedom
Multiple R-squared:  0.0512,Adjusted R-squared:  0.049
F-statistic: 22.6 on 1 and 418 DF,  p-value: 2.78e-06
```

```r
## This only works if `ase` has been loaded
summary_rob(lm(testscore ~ str, data = CASchools))
```

```

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  698.933     10.364   67.44  < 2e-16
str           -2.280      0.519   -4.39  1.1e-05
---
Heteroskadasticity robust standard errors used

Residual standard error: 19 on 418 degrees of freedom
Multiple R-squared:  0.0512, Adjusted R-squared:  0.049
F-statistic: 19.3 on 1 and Inf DF,  p-value: 1.14e-05
```

Difference in means/regression

		testscore	
D	n	mean	sd
0	177	650.00	17.97
1	243	657.18	19.29
All	420	654.16	19.05

$$\bar{Y}_{small} - \bar{Y}_{large} = 657.18 - 650.00$$
$$= 7.18$$

$$SE(\bar{Y}_{small} - \bar{Y}_{large}) = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}$$
$$= 1.83$$

```
summary_rob(lm(testscore ~ D, data = CASchools))


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   650.00       1.35  481.55  < 2e-16
D               7.19       1.83    3.92  8.7e-05
---
Heteroskadasticity robust standard errors used

Residual standard error: 19 on 418 degrees of freedom
Multiple R-squared:  0.0348, Adjusted R-squared:  0.0324
F-statistic: 15.4 on 1 and Inf DF,  p-value: 8.73e-05
```

Some Additional Theoretical Foundations of OLS

Some Additional Theoretical Foundations of OLS (Section 5.5)

We have already learned a very great deal about OLS: (i) OLS is unbiased and consistent; (ii) we have a formula for heteroskedasticity-robust standard errors; (iii) and we can construct confidence intervals and test statistics.

Some Additional Theoretical Foundations of OLS (Section 5.5)

We have already learned a very great deal about OLS: (i) OLS is unbiased and consistent; (ii) we have a formula for heteroskedasticity-robust standard errors; (iii) and we can construct confidence intervals and test statistics.

Also, a very good reason to use OLS is that everyone else does — so by using it, others will understand what you are doing. In effect, OLS is the language of regression analysis, and if you use a different estimator, you will be speaking a different language.

Further questions you may have:

- Is this really a good reason to use OLS? Arent there other estimators that might be better — in particular, ones that might have a smaller variance?

Further questions you may have:

- Is this really a good reason to use OLS? Arent there other estimators that might be better — in particular, ones that might have a smaller variance?

So we will now answer this question but to do so we will need to make some stronger assumptions than the three least squares assumptions already presented.

The Extended Least Squares Assumptions

1. $E(u_i|X_i = x) = 0$, for all x;
2. (X_i, Y_i), $i = 1, \ldots, n$, are $i.i.d.$;
3. large outliers are rare $(E(Y^4) < \infty$, $E(X^4) < \infty)$;
4. u_i is homoskedastic;
5. u_i is $N(0, \sigma_u^2)$.

Efficiency of OLS: The Gauss-Markov Theorem

Gauss-Markov theorem - Part I

Under extended LS assumptions 1-4 (1-3, plus homoskedasticity):

OLS has the smallest variance among all linear estimators of β_1.

Efficiency of OLS:The Gauss-Markov Theorem

Gauss-Markov theorem - Part II

Under extended LS assumptions 1-5 (1-3, plus homoskedasticity and normality):

OLS has the smallest variance among all consistent estimators.

This is a pretty amazing result — it says that, if (in addition to LSA 1-3) the errors are homoskedastic and normally distributed, then OLS is a better choice than any other consistent estimator.

And because an estimator that isn't consistent is a poor choice, this says that OLS really is the best you can do — if extended LS assumptions hold.

Some not-so-good thing about OLS

The foregoing results are impressive, but these results and the OLS estimator have important limitations.

- The GM theorem really isnt that compelling:

Some not-so-good thing about OLS

The foregoing results are impressive, but these results and the OLS estimator have important limitations.

- The GM theorem really isnt that compelling:
    - The condition of homoskedasticity often doesn't hold (homoskedasticity is special)

Some not-so-good thing about OLS

The foregoing results are impressive, but these results and the OLS estimator have important limitations.

- The GM theorem really isnt that compelling:
    - The condition of homoskedasticity often doesn't hold (homoskedasticity is special)
    - The result is only for linear estimators — only a small subset of estimators

Some not-so-good thing about OLS

The foregoing results are impressive, but these results and the OLS estimator have important limitations.

- The GM theorem really isnt that compelling:
  - The condition of homoskedasticity often doesn't hold (homoskedasticity is special)
  - The result is only for linear estimators — only a small subset of estimators
- The strongest optimality result (part II above) requires homoskedastic normal errors - not plausible in applications

Some not-so-good thing about OLS

The foregoing results are impressive, but these results and the OLS estimator have important limitations.

- The GM theorem really isnt that compelling:
    - The condition of homoskedasticity often doesn't hold (homoskedasticity is special)
    - The result is only for linear estimators — only a small subset of estimators
- The strongest optimality result (part II above) requires homoskedastic normal errors - not plausible in applications
- OLS is more sensitive to outliers than some other estimators.

Some not-so-good thing about OLS

The foregoing results are impressive, but these results and the OLS estimator have important limitations.

- The GM theorem really isnt that compelling:
  - The condition of homoskedasticity often doesn't hold (homoskedasticity is special)
  - The result is only for linear estimators — only a small subset of estimators
- The strongest optimality result (part II above) requires homoskedastic normal errors - not plausible in applications
- OLS is more sensitive to outliers than some other estimators.

Some not-so-good thing about OLS

The foregoing results are impressive, but these results and the OLS estimator have important limitations.

- The GM theorem really isnt that compelling:
  - The condition of homoskedasticity often doesn't hold (homoskedasticity is special)
  - The result is only for linear estimators — only a small subset of estimators
- The strongest optimality result (part II above) requires homoskedastic normal errors - not plausible in applications
- OLS is more sensitive to outliers than some other estimators.

In virtually all applied regression analysis, OLS is used and that is what we will do in this course too.