Applied Statistics and Econometrics
Lecture 8

Giuseppe Ragusa
gragusa@luiss.it

March 20, 2017

Wage and education in Italy

- Using the labor force survey we will try to get a sense of the relation between wages and education by estimating a linear model
- We won't be able to turn these estimates into effects because we are not able to control for omitted variables that are unobservables (ability)
- Nevertheless, we will able to lear a great deal about this relation

Italian labor force survey

The Italian Labour Force Survey (Lfs) provides data on labour market variables (employment status, type of work, work experience, job search, etc.), disaggregated by gender, age and territory (up to regional detail on a quarterly base).

```
##    RETRIC ETAM DETIND TISTUD REG SG11 SG16
## 1    1530   50      2     10  10    2    1
## 6    1600   61      2     10  14    2    1
## 7    1500   46      2     10  17    2    1
## 10   2800   43      2      3   1    1    1
## 11   1300   33      2      4   4    1    2
## 12    940   38      1      3   4    2    2
## 16   1700   57      2      5  18    1    1
## 21   2180   32      2      5   1    1    1
## 25   1470   52      2      4  10    1    1
## 26    700   50      2      5  10    2    1
## 45   1800   46      2      5   5    1    1
## 46   1100   42      2      3   5    2    1
## 48   1550   50      2      4   5    1    1
## 49   1250   44      2      3   5    2    1
```

LFS

- `RETRIC` - Net monthly wage
- `ETAM` - Age
- `DETIND` - Temporary/full time worker
- `TISTUD` - Educational attainment
- `SG24b` - Educational attainment (>BA)
- `REG` - Region
- `SG11` - Gender
- `SG16` - Italian Citizen

RETRIC

RETRIC **Retribuzione netta del mese scorso escluse altre mensilità (tredicesima, quattordicesima, ecc.) e voci accessorie non percepite regolarmente tutti i mesi (premi di produttività annuali, arretrati, indennità per missioni, straordinari non abituali, ecc.)**

▪ Fino a 250 euro	250
▪ 260	260
▪ 270	270
▪ -------------	------
▪ -------------	------
▪ -------------	------
▪ 2980	2980
▪ 2990	2990
▪ 3000 euro o più	3000

TISTUD

TISTUD **Titolo di studio a 10 modalità**

▪ Nessun titolo	1
▪ Licenza elementare / Attestato di valutazione finale	2
▪ Licenza media (dall'anno 2007 denominata "Diploma di Istruzione secondaria di I grado") o avviamento professionale (conseguito non oltre all'anno 1965)	3
▪ Diploma di qualifica professionale di scuola secondaria superiore (di II grado) di 2-3 anni che non permette l'iscrizione all'Università /Attestato IFP di qualifica professionale triennale (operatore)/ Diploma professionale IFP di tecnico (quarto anno) (dal 2005)	4
▪ Diploma di maturità / Diploma di istruzione secondaria superiore (di II grado) di 4-5 anni che permette l'iscrizione all'Università/Certificato di specializzazione tecnica superiore IFTS (dal 2000)/ Diploma di tecnico superiore ITS (corsi biennali) (dal 2013)	5
▪ Diploma di Accademia (Belle Arti, Nazionale di arte drammatica, Nazionale di Danza), Istituto superiore Industrie artistiche, Conservatorio di musica statale, Istituto di Musica Pareggiato	6
▪ Diploma universitario di due/tre anni, Scuola diretta a fini speciali, Scuola parauniversitaria	7
▪ Laurea di primo livello (triennale)	8
▪ Laurea specialistica/magistrale biennale	9
▪ Laurea di 4-6 anni: laurea del vecchio ordinamento o laurea specialistica/magistrale a ciclo unico	10

SG24

SG24B. Ha conseguito un titolo di studio post-laurea, post-diploma accademico AFAM o dottorato di ricerca?

- Master universitario di I livello/ Diploma accademico di perfezionamento o Master di I livello/ Diploma accademico di specializzazione di I livello 1 *(passare a SG25)*
- Master universitario di II livello/ Diploma accademico di perfezionamento o Master di II livello/ Diploma accademico di specializzazione di II livello 2 *(passare a SG25)*
- Diploma di specializzazione universitaria 3*(passare a SG25)*
- Dottorato di ricerca/Diploma accademico di formazione alla ricerca AFAM 4*(passare a SG25)*
- Nessuno di questi 5*(passare a SG25)*

REGION

ALLEGATO: REGIONI

Piemonte	*01*	*Marche*	*11*
Valle d'Aosta	*02*	*Lazio*	*12*
Lombardia	*03*	*Abruzzo*	*13*
Trentino Alto Adige	*04*	*Molise*	*14*
Veneto	*05*	*Campania*	*15*
Friuli Venezia Giulia	*06*	*Puglia*	*16*
Liguria	*07*	*Basilicata*	*17*
Emilia Romagna	*08*	*Calabria*	*18*
Toscana	*09*	*Sicilia*	*19*
Umbria	*10*	*Sardegna*	*20*

GENDER

SG11. Sesso del componente

- Maschio 1
- Femmina 2

SG16. Cittadinanza italiana

- Sì 1
- No 2 *(passare a SG17)*

Recode variables

```r
library(dplyr)
lfs <- lfs %>% mutate(female = ifelse(SG11 == 2, 1, 0))
lfs <- lfs %>% mutate(citizen = ifelse(SG16 == 1, 1, 0))

library(car)
lfs$educ <- car::recode(lfs$TISTUD, "10=19;9=19;8=16;7=15;6=16;5=13;4=10;3=13;2=5;1=0")
lfs$SG24B <- car::recode(lfs$SG24B, "NA=0")
lfs$educ[lfs$SG24B == 4] <- 21
lfs$educ[lfs$SG24B == 3] <- lfs$educ[lfs$SG24B == 3] + 1
```

Wage and years of education

```
lm_1 <- lm(RETRIC ~ educ, data = lfs)
summary_rob(lm_1)
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   628.35      15.45    40.7   <2e-16
## educ           50.44       1.18    42.8   <2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 500 on 26125 degrees of freedom
## Multiple R-squared:  0.0827, Adjusted R-squared:  0.0827
## F-statistic: 1.83e+03 on 1 and Inf DF,  p-value: <2e-16
```

Gender Gap

Gender gap refers to systematic differences in the outcomes that men and women achieve in the labor market.
A "vanilla" gender gap can be estimated

$$wage = \beta_0 + \beta_1 female + u_i$$

```
lm_2 <- lm(RETRIC ~ female, data = lfs)
summary_rob(lm_2)

##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1444.54       4.42     327   <2e-16
## female       -291.36       6.19     -47   <2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 502 on 26125 degrees of freedom
## Multiple R-squared:  0.0775, Adjusted R-squared:  0.0775
## F-statistic: 2.21e+03 on 1 and Inf DF,  p-value: <2e-16
```

Wage-(potential) experience relation

Potential experience is the maximum years of experience of the individual in the job market. Actual experience is usually unavailable. Potential experience is defined

$X =$ age-years at graduation.

In the `lfs` we do not observe years at graduation, so we will use *age* to proxy for experience.

```
lm_3 <- lm(RETRIC ~ ETAM, data = lfs)
summary_rob(lm_3)

##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  806.977     12.119    66.6   <2e-16
## ETAM          11.367      0.285    39.9   <2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 508 on 26125 degrees of freedom
## Multiple R-squared:  0.0564, Adjusted R-squared:  0.0564
## F-statistic: 1.59e+03 on 1 and Inf DF,  p-value: <2e-16
```

Multivariate regression

$$wage = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 female + u_i$$

```
summary_rob(lm(RETRIC ~ educ + ETAM + female, data = lfs))
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   127.026     19.143    6.64  3.2e-11
## educ           57.788      1.079   53.56  < 2e-16
## ETAM           12.758      0.258   49.48  < 2e-16
## female       -334.885      5.660  -59.17  < 2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 453 on 26123 degrees of freedom
## Multiple R-squared:  0.247, Adjusted R-squared:  0.247
## F-statistic: 6.99e+03 on 3 and Inf DF,  p-value: <2e-16
```

Regional differences

Consider the variable `RIP3`

RIP3	**Ripartizione geografica in 3 classi**		
	▪	Nord	1
	▪	Centro	2
	▪	Mezzogiorno	3

How we asses the differences in wages among these parts of the country?

We could think of running the following regression:

$$wage_i = \beta_0 + \beta_1 South_i + \beta_2 Center_i + \beta_3 North_i + u_i$$

where

- $South_i$ is a dummy taking value 1 if individual i resides in the southernmost part of Italy
- $Center_i$ is a dummy taking value 1 if individual i resides in the south of Italy
- $North_i$ is a dummy taking value 1 if individual i resides in the northernmost part of Italy

We could....but it is not probably a good idea....

Dummy variables

$$wage_i = \beta_0 + \beta_1 South_i + \beta_2 Center_i + \beta_3 North_i + u_i$$

What is the interpretation of β_1 in this regression?

$$E[wage_i | South_i = 1, Center_i = 0, North_i = 0] = \beta_0 + \beta_1$$

$$E[wage_i | South_i = 0, Center_i = 0, North_i = 0] = \beta_0$$

.... but at least one of $South_i$, $Center_i$, and $North_i$ must be 1 because the categories are exclusive.

Dummy Variable Trap

If you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive – that is, there are multiple categories and every observation falls in one and only one category and include all these dummy variables and a constant, you will have perfect multicollinearity – this is sometimes called the dummy variable trap.

Dummy variables: omit a category

We consider instead

$$wage_i = \beta_0 + \beta_1 South_i + \beta_2 Center_i + u_i$$

What is the interpretation of β_1 in this regression?

$$E[wage_i | South_i = 1, Center_i = 0] = \beta_0 + \beta_1$$

$$E[wage_i | South_i = 0, Center_i = 0] = \beta_0$$

Now this is ok, because when $South_i = 0$ and $Center_i = 0$ it means that i resides in the northernmost region. Thus,

$$\beta_1 = \underbrace{E[wage_i | South_i = 1, Center_i = 0]}_{\text{avg. wage in the south}} - \underbrace{E[wage_i | South_i = 0, Center_i = 0]}_{\text{avg. wage in the north}}$$

β_1 is the average difference in wage between workers in the south and workers in the north.

Dummy variables: omit the intercept

We consider instead

$$wage_i = \beta_1 South_i + \beta_2 Center_i + \beta_3 North_i + u_i$$

What is the interpretation of β_1 in this regression?

$$E[wage_i | South_i = 1, Center_i = 0, North_i = 0] = \beta_1$$

$$E[wage_i | South_i = 0, Center_i = 1, North_i = 0] = \beta_2$$

$$E[wage_i | South_i = 0, Center_i = 0, North_i = 1] = \beta_3$$

Creating dummy

Let's create the dummy variables.

```
lfs <- lfs %>% mutate(South = ifelse(RIP3 == 3, 1, 0), North = ifelse(RIP3 == 1,
    1, 0), Center = ifelse(RIP3 == 2, 1, 0))
head(lfs[c("South", "Center", "North")])
```

```
##   South Center North
## 1     0      1     0
## 2     1      0     0
## 3     1      0     0
## 4     0      0     1
## 5     0      0     1
## 6     0      0     1
```

Geographic wage differentials I

```
summary_rob(lm(RETRIC ~ South + Center, data = lfs))
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1349.79       4.26  316.75  < 2e-16
## South        -133.64       8.00  -16.71  < 2e-16
## Center        -65.62       8.25   -7.95  1.8e-15
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 520 on 26124 degrees of freedom
## Multiple R-squared:  0.0106, Adjusted R-squared:  0.0105
## F-statistic:  291 on 2 and Inf DF,  p-value: <2e-16
```

Geographic wage differentials II

```r
summary_rob(lm(RETRIC ~ South + Center + North - 1, data = lfs))
```

```
##
## Coefficients:
##        Estimate Std. Error z value Pr(>|z|)
## South   1216.15       6.77     180   <2e-16
## Center  1284.17       7.07     182   <2e-16
## North   1349.79       4.26     317   <2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 520 on 26124 degrees of freedom
## Multiple R-squared:  0.864, Adjusted R-squared:  0.864
## F-statistic: 1.66e+05 on 2 and Inf DF,  p-value: <2e-16
```

Geographic wage differentials III

If all the dummy variables are included in the regression Rwill automatically drop one in order to be able to proceed.

```
lm(RETRIC ~ South + Center + North, data = lfs)
```

```
##
## Call:
## lm(formula = RETRIC ~ South + Center + North, data = lfs)
##
## Coefficients:
## (Intercept)        South       Center        North
##      1349.8       -133.6        -65.6           NA
```

Which variable should be omitted?

The conclusions we draw from the regression do not change with the specific variable that is omitted

$$RETRIC = \underset{(6.8)}{1216.2} + \underset{(9.8)}{68}\,Center + \underset{(8)}{133.6}\,North$$

$$RETRIC = \underset{(7.1)}{1284.2} - \underset{(9.8)}{68}\,South + \underset{(8.3)}{65.6}\,North$$

$$RETRIC = \underset{(4.3)}{1349.8} - \underset{(8)}{133.6}\,South - \underset{(8.3)}{65.6}\,Center$$

$$RETRIC = +\underset{(6.8)}{1216.2}\,South + \underset{(7.1)}{1284.2}\,Center + \underset{(4.3)}{1349.8}\,North$$

Exclusive dummies in R

Dealing with categorical variables using their dummy representation is so common that Rhas a mechanism to deal with it. This mechanism is based on factor.

```
lfs$RIP <- factor(lfs$RIP3, level = c(1, 2, 3), labels = c("North", "Center", "South"))

table(lfs$RIP)

##
##  North Center  South
##  14738   5845   5544

head(lfs$RIP)

## [1] Center South  South  North  North  North
## Levels: North Center South
```

Exclusive dummies in R

```r
summary_rob(lm(RETRIC ~ RIP, data = lfs))
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1349.79       4.26  316.75  < 2e-16
## Center        -65.62       8.25   -7.95  1.8e-15
## South        -133.64       8.00  -16.71  < 2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 520 on 26124 degrees of freedom
## Multiple R-squared:  0.0106, Adjusted R-squared:  0.0105
## F-statistic:  291 on 2 and Inf DF,  p-value: <2e-16
```

Wage and education and controls

```r
lm_full <- lm(RETRIC ~ educ + ETAM + female + RIP, data = lfs)
summary_rob(lm_full)
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  167.867     19.123    8.78   <2e-16
## educ          58.227      1.076   54.11   <2e-16
## ETAM          13.100      0.256   51.18   <2e-16
## female      -342.400      5.610  -61.03   <2e-16
## Center       -95.489      7.076  -13.50   <2e-16
## South       -173.778      6.831  -25.44   <2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 448 on 26121 degrees of freedom
## Multiple R-squared:  0.266, Adjusted R-squared:  0.265
## F-statistic: 7.7e+03 on 5 and Inf DF,  p-value: <2e-16
```

Interpreting the coefficient on geographic areas

$$RETRIC = \underset{(19.12)}{167.87} + \underset{(1.08)}{58.23}\, educ + \underset{(0.26)}{13.1}\, ETAM - \underset{(5.61)}{342.4}\, female - \underset{(7.08)}{95.49}\, Center - \underset{(6.83)}{173.78}\, South$$

In the above regression, the interpretation of the estimated coefficients on `Center` and `South` is the following:

- Individuals living in the south make on average **173 euro less** than individual living in the north, everything else being equal.
- Individuals living in the center make **95 euro less** than individual living in the north, everything else being equal.

Interpreting the coefficient on geographic areas

What does it mean to test for the coefficient on *South* being equal to zero?

It means testing whether, everything else being equal, wages of individuals living in the south **are not different** on average to those in the *North*.

Testing geographic differences in wages

$$RETRIC = \beta_0 + \beta_1 educ + \beta_2 ETAM + \beta_3 female + \beta_4 Center + \beta_5 South + u_i$$

$$RETRIC = \underset{(19.12)}{167.87} + \underset{(1.08)}{58.23}\, educ + \underset{(0.26)}{13.1}\, ETAM - \underset{(5.61)}{342.4}\, female - \underset{(7.08)}{95.49}\, Center - \underset{(6.83)}{173.78}\, South$$

Testing equality btw North and South

$$H_0 : \beta_4 = 0, \quad vs. \quad H_1 : \beta_4 \neq 0$$

$$t = \frac{-95.49}{7.08} = -13.49$$

Testing equality btw North and Center

$$H_0 : \beta_5 = 0, \quad vs. \quad H_1 : \beta_5 \neq 0$$

$$t = \frac{-173.78}{6.83} = -25.44$$

Testing geographic differences in wages

The t-tests on the coefficients imply:

- Wages in the center are statistically different from those in the north.
- Wages in the center are statistically different from those in the north.

Does these conclusions imply that wages **in the center and in the south** are statistically different from those in the north.

Short answer: no!

Long answer: We need to consider the following null hypothesis:

$$H_0 : \beta_4 = 0 \ AND \ \beta_5 = 0$$

Tests of joint hypotheses, ctd.

- $$H_0 : \beta_1 = 0, \ \textbf{and } \beta_2 = 0$$
  $$vs. \ H_1 : \text{either } \beta_1 \neq 0 \text{ or } \beta 2 {\neq} 0 \text{ or both}$$
- A joint hypothesis specifies a value for two or more coefficients, that is, it imposes a restriction on two or more coefficients.
- In general, a joint hypothesis will involve q restrictions. In the example above, q = 2, and the two restrictions are $\beta_1 = 0$ and $\beta_2 = 0$.
- A "common sense" idea is to reject if either of the individual t-statistics exceeds 1.96 in absolute value.
- But this "one at a time" test **isn't valid**: the resulting test rejects too often under the null hypothesis (more than 5%)!

Why can't we just test the coefficients one at a time?

Because the rejection rate under the null isn't 5%. We'll calculate the probability of incorrectly rejecting the null using the "common sense" test based on the two individual t-statistics. To simplify the calculation, suppose that and are independently distributed (this isn't true in general – just in this example). Let t_1 and t_2 be the t-statistics:

$$t_1 = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}, \quad t_2 = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)},$$

The "one at time" test is:

$$\text{reject } H_0 : \beta_1 = \beta_2 = 0 \text{ if } |t_1| > 1.96 \text{ and/or } |t_2| > 1.96$$

What is the probability that this "one at a time" test rejects H_0, when H_0 is actually true? (It should be 5%.)

Suppose t_1 and t_2 are independent (for this example)

The probability of incorrectly rejecting the null hypothesis using the "one at a time" test

$$
\begin{aligned}
&= \Pr_{H_0}[|t_1| > 1.96 \text{ and/or } |t_2| > 1.96] \\
&= 1 - \Pr_{H_0}[|t_1| \leq 1.96 \text{ and } |t_2| \leq 1.96] \\
&= 1 - \underbrace{\Pr_{H_0}[|t_1| \leq 1.96] \times \Pr_{H_0}[|t_2| \leq 1.96]}_{\text{(because } t_1 \text{ and } t_2 \text{ are independent by assumption)}} \\
&= 1 - (.95)^2 = 0.0975 = 9.75\%
\end{aligned}
$$

Which is **not** the desired 5%!!

- In fact, its size depends on the correlation between t_1 and t_2 (and thus on the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$).

Two solutions:

- Use a different critical value in this procedure – not 1.96 (this is the "Bonferroni method – see SW App. 7.1) (this method is rarely used in practice however)
- Use a different test statistic designed to test both β_1 and β_2 at once: the Wald test (this is common practice)

The Wald statistic

- The F-statistic tests all parts of a joint hypothesis at once.
- The formula for the special case of the joint hypothesis

$$H_0 : \beta_1 = \beta_{1,0} \textbf{ and } \beta_2 = \beta_{2,0}$$

is

$$W = n \times \begin{pmatrix} \hat{\beta}_1 - \beta_{1,0} \\ \hat{\beta}_2 - \beta_{2,0} \end{pmatrix}' \begin{bmatrix} \hat{\sigma}^2_{\hat{\beta}_1} & \hat{\sigma}_{\hat{\beta}_1,\hat{\beta}_2} \\ \hat{\sigma}_{\hat{\beta}_2,\hat{\beta}_1} & \hat{\sigma}^2_{\hat{\beta}_2} \end{bmatrix}^{-1} \begin{pmatrix} \hat{\beta}_1 - \beta_{1,0} \\ \hat{\beta}_2 - \beta_{2,0} \end{pmatrix}$$

- Reject when W is large (how large?)

Large-sample distribution of the Wald-statistic

In large samples the formula becomes

$$W \xrightarrow{d} \chi^2_q$$

- The large-sample distribution of the Wald statistic is the distribution of the average of two independently distributed squared standard normal random variables.

The chi-squared distribution

The chi-squared distribution with q degrees of freedom χ^2_q is defined to be the distribution of the sum of q independent squared standard normal random variables. Formally, if

$$Z_1 \xrightarrow{d} N(0,1),\ Z_2 \xrightarrow{d} N(0,1), \dots, Z_q \xrightarrow{d} N(0,1),$$

then

$$Z_1^2 + Z_2^2 + \dots + Z_q^2 \xrightarrow{d} \chi^2_q.$$

Selected large-sample critical values of χ^2_1

q	5% crit. val.	10% crit. val.
1.00	3.84	2.71
2.00	5.99	4.61
3.00	7.81	6.25
4.00	9.49	7.78
5.00	11.07	9.24

Implementation in R

Example

Test the joint hypothesis that the population coefficients on *str* and expenditures per pupil (*expenditure*) are both zero, against the alternative that at least one of the population coefficients is nonzero.

$$testscore_i = \beta_0 + \beta_1 str_i + \beta_2 expenditure_i + \beta_3 english_i + u_i$$

1. Estimate the model using `lm(...)`
2. Use `wald_test(...)` to conduct the test (this function is in `ase`)

Implementation in R: California Schools

```r
lm_cas <- lm(testscore ~ str + expenditure +
    english, data = CASchools)
summary_rob(lm_cas)



Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 649.57795   15.45834   42.02   <2e-16
str          -0.28640    0.48207   -0.59    0.552
expenditure   0.00387    0.00158    2.45    0.014
english      -0.65602    0.03178  -20.64   <2e-16
---
Heteroskadasticity robust standard errors used

Residual standard error: 14.4 on 416 degrees of freedom
Multiple R-squared:  0.437, Adjusted R-squared:  0.433
F-statistic:  442 on 3 and Inf DF,  p-value: <2e-16
```

```r
wald_test(lm_cas, testcoef = c("str", "expenditure"))

Wald test

Null hypothesis:
str = 0
expenditure = 0

 q     W   pvalue
 2 10.87 0.004367
```

More on Wald statistic

- The general for of the Wald statistics can be expressed using matrix algebra. Let

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V)$$

where

$$V = \begin{pmatrix} \sigma^2_{\hat{\beta}_0} & \sigma_{\hat{\beta}_0\hat{\beta}_1} & \cdots & \sigma_{\hat{\beta}_0\hat{\beta}_k} \\ \sigma_{\hat{\beta}_1\hat{\beta}_0} & \sigma^2_{\hat{\beta}_1} & \cdots & \sigma_{\hat{\beta}_1\hat{\beta}_k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\hat{\beta}_k\hat{\beta}_0} & \sigma_{\hat{\beta}_k\hat{\beta}_1} & \cdots & \sigma^2_{\hat{\beta}_k} \end{pmatrix}$$

- Let $\hat{V}$ denote an estimator of V

$$\hat{V} = \begin{pmatrix} \hat{\sigma}^2_{\hat{\beta}_0} & \hat{\sigma}_{\hat{\beta}_0\hat{\beta}_1} & \cdots & \hat{\sigma}_{\hat{\beta}_0\hat{\beta}_k} \\ \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_0} & \hat{\sigma}^2_{\hat{\beta}_1} & \cdots & \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_k} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{\hat{\beta}_k\hat{\beta}_0} & \hat{\sigma}_{\hat{\beta}_k\hat{\beta}_1} & \cdots & \hat{\sigma}^2_{\hat{\beta}_k} \end{pmatrix}$$

More on Wald statistic

Suppose we want to test the following null hypothesis

$$H_0 : \beta_1 = \beta_{1,0} \text{ and } \beta_3 = \beta_{3,0}$$

This null can be written in terms of matrix and vector. Let

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & \cdots & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

Thus, the null hypothesis can be written as

$$H_0 : R\beta = \bar{\beta}$$

More on Wald statistic, ctd

Using this notation, the Wald statistic can be written as

$$W = n \times (\hat{\beta}' R' - \bar{\beta}') \left[R \hat{V} R' \right]^{-1} (R\hat{\beta} - \bar{\beta})$$

where

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

More on Wald statistic, ctd

and

$$R\hat{V}R' = \begin{pmatrix} \hat{\sigma}^2_{\hat{\beta}_1} & \hat{\sigma}_{\hat{\beta}_1,\hat{\beta}_3} \\ \hat{\sigma}_{\hat{\beta}_3,\hat{\beta}_1} & \hat{\sigma}^2_{\hat{\beta}_3} \end{pmatrix}$$

Thus,

$$\begin{aligned} W &= n \times (\hat{\beta} - \bar{\beta})' R' \left[R\hat{V}R' \right]^{-1} R(\hat{\beta} - \bar{\beta}) \\ &= n \times \begin{pmatrix} \hat{\beta}_1 - \beta_{1,0} \\ \hat{\beta}_3 - \beta_{3,0} \end{pmatrix}' \begin{pmatrix} \hat{\sigma}^2_{\hat{\beta}_1} & \hat{\sigma}_{\hat{\beta}_1,\hat{\beta}_3} \\ \hat{\sigma}_{\hat{\beta}_3,\hat{\beta}_1} & \hat{\sigma}^2_{\hat{\beta}_3} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\beta}_1 - \beta_{1,0} \\ \hat{\beta}_3 - \beta_{3,0} \end{pmatrix} \end{aligned}$$

More on Wald statistic, ctd

- Notice that, for instance,

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}/n}$$

- Thus, when $q = 1$, the wald test is the square of the $t - statistics$

$$\begin{aligned} W &= (\hat{\beta}_1 - \beta_{1,0}) \left[\hat{\sigma}^2_{\hat{\beta}_1}/n\right]^{-1} (\hat{\beta}_1 - \beta_{1,0}) \\ &= (\hat{\beta}_1 - \beta_{1,0}) \left[\hat{\sigma}^2_{\hat{\beta}_1}/n\right]^{-1/2} \left[\hat{\sigma}^2_{\hat{\beta}_1}/n\right]^{-1/2} (\hat{\beta}_1 - \beta_{1,0}) \\ &= \left[\frac{(\hat{\beta}_1 - \beta_{1,0})}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}/n}}\right]^2 \end{aligned}$$

- So, the Wald statistic is a generalization of the t-test

Wald statistic and F-statistic

- The book (cf. Section 7.2) discusses the $F-$Statistics.
- The $F-$Statistics of Stock and Watson is proportional to our Wald-statistic

$$W = q \times F$$

- They are really the same thing, but you need to be careful in adjusting the critical value

$$W \xrightarrow{d} \chi^2_q, \quad F \xrightarrow{d} \chi^2_q / q$$

- The Wald-statistic formula is very general

$$W = n \times (\hat{\beta}' R' - \bar{\beta}') \left[R \hat{V} R' \right]^{-1} (R\hat{\beta} - \bar{\beta})$$

For instance, if we are willing to assume homoskedasticity the only thing we have to change is $\hat{V}$, the estimated variance

Wald-statistic with homoskedasticity

In `R`, $\hat{V}/n$ can be obtained by using `vcov(object)` (homoskedastic) and `vcov(object)` from the package `sandwich` (heteroskedastic robust) where `object` is an object of class `lm`

```
vcovHC(lm_cas)
```

```
            (Intercept)        str expenditure    english
(Intercept)   245.50574 -6.8268127  -2.136e-02  8.529e-02
str            -6.82681  0.2376689   4.116e-04 -2.531e-03
expenditure    -0.02136  0.0004116   2.584e-06 -1.065e-05
english         0.08529 -0.0025306  -1.065e-05  1.031e-03
```

```
vcovHC(lm_cas)
```

```
            (Intercept)        str expenditure    english
(Intercept)   245.50574 -6.8268127  -2.136e-02  8.529e-02
str            -6.82681  0.2376689   4.116e-04 -2.531e-03
expenditure    -0.02136  0.0004116   2.584e-06 -1.065e-05
english         0.08529 -0.0025306  -1.065e-05  1.031e-03
```

Wald-statistic with homoskedasticity and heteroskedasticity

Heteroskedastic robust

```r
wald_test(lm_cas, testcoef = c("str", "expenditure"))
```

```
## Wald test
##
## Null hypothesis:
## str = 0
## expenditure = 0
##
##  q     W   pvalue
##  2 10.87 0.004367
```

Homoskedastic only

```r
wald_test(lm_cas, testcoef = c("str", "expenditure"),
    vcov = vcov)
```

```
## Wald test
##
## Null hypothesis:
## str = 0
## expenditure = 0
##
##  q     W    pvalue
##  2 16.02 0.0003321
```

Wald statistic with homoskedasticity

- When the errors are homoskedastic, there is a simple formula for computing the "**homoskedasticity-only**" Wald-statistic:
    1. Run two regressions, one under the null hypothesis (the "restricted" regression) and one under the alternative hypothesis (the "unrestricted" regression).
    2. Compare the fits of the regressions – the R^2s – if the "unrestricted" model fits sufficiently better, reject the null

The "restricted" and "unrestricted" regressions

are the coefficients on str and expenditure zero?

Unrestricted population regression (under H_1):

$$testscore_i = \beta_0 + \beta_1 str_i + \beta_2 expn_i + \beta_3 english_i + u_i$$

Restricted population regression (that is, under H_0):

$$testscore_i = \beta_0 + \beta_3 english_i + u_i$$

- The number of restrictions under H_0 is $q = 2$ (why?).
- The fit will be better (R^2 will be higher) in the unrestricted regression (why?)
- By how much must the R^2 increase for the coefficients on *expn* and *english* to be judged statistically significant? There is a formula for this...

Simple formula for the homoskedasticity-only Wald-statistic:

$$W = q \times \frac{(R^2_{\text{unrestricted}} - R^2_{\text{restricted}})/q}{(1 - R^2_{\text{unrestricted}})/(n - k_{\text{unrestricted}} - 1)}$$

where

1. $R^2_{\text{unrestricted}}$: the R^2 for the unrestricted regression
2. $R^2_{\text{restricted}}$: the R^2 for the restricted regression
3. q: the number of restriction *under the null hypothesis*
4. $k_{\text{unrestricted}}$ the number of regressors in the unrestricted regression

The bigger the difference between the restricted and unrestricted R^2s – the greater the improvement in fit by adding the variables in question – the larger is the homoskedasticity-only Wald-statistic.

Simple formula for Wald-statistic.

- **Unrestricted** model

$$testscore = 692.69 - 1.351\, str - 0.13\, expn - 0.67\, english, \quad R^2 = 0.4366$$

- **Restricted** model

$$testscore = 664.739 - 0.671\, english, \quad R^2 = 0.4149$$

$$\begin{aligned} W &= q \times \frac{(R^2_{\text{unrestricted}} - R^2_{\text{restricted}})/q}{(1 - R^2_{\text{unrestricted}})/(n - k_{\text{unrestricted}} - 1)} \\ &= 2 \times \frac{0.0217/2}{1 - 0.4366/(420 - 3 - 1)} = 16.02 \end{aligned}$$

Summary: testing joint hypotheses

- The "one at a time" approach of rejecting if either of the t-statistics exceeds 1.96 rejects more than 5% of the time under the null (the size exceeds the desired significance level)
- The heteroskedasticity-robust Wald-statistic can be calculated using `R` (`wald_test` command); this tests all `q` restrictions at once.
- For n large, the Wald-statistic is distributed χ^2_q
- The homoskedasticity-only Wald-statistic is important historically (and thus in practice), and can help intuition, but isn't valid when there is heteroskedasticity