Regression analysis

Ozan Bakış[1]

[1] Bahcesehir University, Department of Economics and BETAM

Outline

1. Regression analysis: introduction
   - Motivation for regression
   - Regression analysis
   - Applications

2. Regression analysis: further issues

Regression as comparison-of-means I

- Say we want to knowwhether women earn less compared to men in USA. One easy way is to compare mean hourly wages using **`wage1`** data, (a sample of 1976 Current Population Survey).

```
url = "https://github.com/obakis/econ_data/raw/master/wage1.rds"
download.file(url, "wage1.rds", mode ="wb")
wage1 = readRDS("wage1.rds")
View(wage1)
```

```
# average wage for men and women.
aggregate(wage ~ female, data = wage1, FUN = mean )
```

```
##   female wage
## 1      0 7.10
## 2      1 4.59
```

```
7.10-4.59
```

```
## [1] 2.51
```

- Even if there is a difference of 2.51 dollars between average hourly wage of women and men, we do not know whether this difference is statistically significant. Is the difference statistically significant?

Regression as comparison-of-means II

- This is why it is better to use a simple regression of the following form

$$wage = \beta_0 + \delta_0 female + u$$

where $female = 0$ for male workers and $female = 1$ for female workers. δ_0 is is the difference in average hourly wage between females and males.

```
reg = lm(wage ~ female, data=wage1)
 ### look at the summary to determine whether the diff. is significant
 summary(reg)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     7.10      0.210   33.81 8.97e-134
## female         -2.51      0.303   -8.28  1.04e-15
```

```
 ### To get back "wage level" for women
 b=coef(reg)
 b
```

```
## (Intercept)      female
##        7.10       -2.51
```

```
 ## male wage = beta_0
 b[1]
```

Regression as comparison-of-means III

```
## (Intercept)
##         7.1
```

```
## female effect = beta_0 + delta_0
b[1]+b[2]
```

```
## (Intercept)
##        4.59
```

A simple regression model as above can be used for ”comparison-of-means” test as above.

- While **`aggregate`** (tables of mean wages) focus on wage levels, **`lm`** (regression) focus on wage differences.

Regression as a tool for ”controlling for” I

- When we estimate the following model
$$wage = \beta_0 + \delta_0 female + u$$
What does a significant δ_0 signify? It represents (total) gender wage gap, for whatever reason (including education).

- Can we say that being female causes lower wages?Or does the above mean that there is a discrimination against women?

- For this to be causal we would need the zero conditional mean assumption: $E(u|female) = 0$ [or equivalently $Cov(female, u) = 0$]. Once we assume this, we can write
$$\delta_0 = E(wage|female = 1) - E(wage|female = 0)$$
so that the difference in hourly wages (δ_0) is only due to gender itself. All other characteristics are assumed to be the same for both genders.

Regression as a tool for ”controlling for” II

- In reality, we doubt that the zero conditional mean assumption holds. There are various factors that differ between women and men. Think about average education, experience, job quality etc. All these factors are likely to affect wages.

- If these factors differ between men and women, then our estimate for δ_0 will be biased. In other words δ_0 will represent both pure gender gap and other effects.

- To decide whether there is discrimination we need to compare men and women with the same observable characteristics. Two problems. First, usually we do not control data collection process and second in real data there are differences especially regarding work experience and education (for older generations).

- So, what to do? Can we change the data?

Regression as a tool for ”controlling for” III

- This is where multiple regression ”controls for” (holds fixed) observable characteristics as if we compare men and women with the same characteristics. A first factor to consider is education.

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

- Now, education is in the model. And the zero conditional mean assumption: $E(u|female, educ) = 0$. Assuming it holds,we can write

$$\delta_0 = E(wage|female = 1, educ) - E(wage|female = 0, educ)$$

δ_0 now represents gender gap in wage for individuals with the same education level. Alternatively, we can sayδ_0 now represents gender gap in wage after correction for differences in education.

Regression as a tool for ”controlling for” IV

- Even if in real life average education levels are not the same the above regression allows us to compute the difference in hourly wages that would occur if men and women had the same level of education on average. This is what controlling for means.

```
 reg = lm(wage ~ female+educ, data=wage1)
summary(reg)$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.623     0.6725   0.926 3.55e-01
## female        -2.273     0.2790  -8.147 2.76e-15
## educ           0.506     0.0504  10.051 7.56e-22
```

As we saw above, once we control for education the difference in hourly wage between females and males is relatively lower. How do you interpret this?Another factor is experience

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + u$$

```
 reg = lm(wage ~ female+educ+exper, data=wage1)
summary(reg)$coef
```

Regression as a tool for ”controlling for” V

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.7345     0.7536   -2.30 2.18e-02
## female       -2.1555     0.2703   -7.97 9.74e-15
## educ          0.6026     0.0511   11.79 1.33e-28
## exper         0.0642     0.0104    6.18 1.32e-09
```

δ_0 is the difference in hourly wage between females and males who have the same amount of education and experience. This is getting lower. How do you interpret this?

Multiple regression in R I

- linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + u_i, \quad i = 1, ..., n.$$

  Application: estimation of wage equation using `hls2011`, a random sample of 762 observations from Turkish Household Labor Force Survey, 2011.

- Get the data:

```
f_url = "https://github.com/obakis/econ_data/raw/master/hls2011.rds"
download.file(url = f_url, destfile = "hls2011.rds", mode="wb")
hls = readRDS("hls2011.rds")
head(hls,3)
```

- A regression example in R:

```
reg = lm(log(hwage) ~ educ + female + emp_sect, data=hls)
summary(reg)
```

Multiple regression in R II

```
##
## Call:
## lm(formula = log(hwage) ~ educ + female + emp_sect, data = hls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0281 -0.2906 -0.0164  0.2619  2.3438
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.76210    0.15824    4.82  1.8e-06 ***
## educ         0.05854    0.00434   13.48  < 2e-16 ***
## female      -0.09318    0.04175   -2.23    0.026 *
## emp_sectpriv  0.08211    0.15739    0.52    0.602
## emp_sectpub  0.82967    0.16168    5.13  3.7e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.463 on 757 degrees of freedom
## Multiple R-squared:  0.536,Adjusted R-squared:  0.534
## F-statistic:  219 on 4 and 757 DF,  p-value: <2e-16
```

Multiple regression in R III

- ”ceteris paribus” interpretation...
- Note that `emp_sect` is a `factor` variable with 3 levels. In R, categorical and ordered categorical (ordinal) variables are called `factors`. Each possible value of a categorical variable is called a level. In a regression a set of dummy variables will be automatically created by R. More precisely, if we have n groups/levels, $n - 1$ dummy variables will be created.
- Operators `+, -, :, *, /, ^` have special meanings in a `formula` object. To ensure arithmetic meaning, we need either to protect by insulation in a function, e.g.,`log(x1 * x2)` or to use `I()` function.
- Generic functions related to `lm` object (See `help(lm)` for details):

Factor variables I

- The `lm()` command, relies on `model.matrix()` for the creation of dummy variables.

```
dummy <- factor(LETTERS[1:4])
model.matrix( ~ dummy)
```

```
##   (Intercept) dummyB dummyC dummyD
## 1           1      0      0      0
## 2           1      1      0      0
## 3           1      0      1      0
## 4           1      0      0      1
## attr(,"assign")
## [1] 0 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$dummy
## [1] "contr.treatment"
```

- To change the base level of a factor variable (ex. ”region” variable) we can use `relevel` function

```
table(hls$emp_sect)
```

Factor variables II

```
## 
## other  priv   pub 
##     9   557   196
```

```r
levels(hls$emp_sect)
```

```
## [1] "other" "priv"  "pub"
```

```r
contrasts(hls$emp_sect) #other is base level
```

```
##       priv pub
## other    0   0
## priv     1   0
## pub      0   1
```

```r
coef(summary(reg))
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.7621    0.15824   4.816 1.77e-06
## educ           0.0585    0.00434  13.484 2.71e-37
## female        -0.0932    0.04175  -2.232 2.59e-02
## emp_sectpriv   0.0821    0.15739   0.522 6.02e-01
## emp_sectpub    0.8297    0.16168   5.131 3.66e-07
```

Factor variables III

```r
hls$emp_sect <- relevel(hls$emp_sect, ref = "pub")
reg_upd <- update(reg, formula = . ~ .) ## we change nothing here!
coef(summary(reg_upd))
```

```
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       1.5918    0.06375   24.97 7.62e-101
## educ              0.0585    0.00434   13.48  2.71e-37
## female           -0.0932    0.04175   -2.23  2.59e-02
## emp_sectother    -0.8297    0.16168   -5.13  3.66e-07
## emp_sectpriv     -0.7476    0.04327  -17.28  1.28e-56
```

Interactions I

- We already saw that the model

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \cdots + u$$

controls for education when analyzing gender wage gap.

- But this model assumes that returns to education are the same for men and women. What if not? Assume that return to education is different

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \times educ + \cdots + u$$

- But now, what is the interpretation of δ_0? It is the wage gap between men and women who has no education $educ = 0$ because the partial effect of female is given by:

$$\frac{\Delta \widehat{wage}}{\Delta female} = \hat{\delta}_0 + \hat{\delta}_1 \, educ$$

Interactions II

- Usually this is not interesting because the average worker has more than zero years of schooling. The general practice is to calculate the wage gap for men and women who have mean education level.

- Even if in the data education levels are not the same this is interpreted as the wage gap that would be observed if there was no difference in terms of education between men and women. We can calculate
$$\left.\frac{\Delta\widehat{wage}}{\Delta female}\right|_{educ=\mu_{educ}} = \hat{\delta}_0 + \hat{\delta}_1\,\mu_{educ}$$
but we cannot know whether the above is significant or not easily, because we lack a t statistic for this expression.

- A second way is to reparameterize the model so that the coefficients on the original variables have an interesting meaning.
$$\widehat{wage} = \hat{\alpha}_0 + \hat{\alpha}_1\,female + \hat{\alpha}_2\,educ + \hat{\alpha}_3 female \times (educ - \mu_{educ}) + \ldots$$
where μ_{educ} is the average years of schooling in the sample.

Interactions III

- The partial effect of female is given by

$$\frac{\Delta\widehat{\mathrm{wage}}}{\Delta\mathrm{female}} = \hat{\alpha}_1 + \hat{\alpha}_3\left(\mathrm{educ} - \mu_{\mathrm{educ}}\right)$$

  Evaluating this at the mean value of education ($\mathrm{educ} = \mu_{\mathrm{educ}}$) we get

$$\left.\frac{\Delta\widehat{\mathrm{wage}}}{\Delta\mathrm{female}}\right|_{\mathrm{educ}=\mu_{\mathrm{educ}}} = \hat{\alpha}_1 = \hat{\delta}_0 + \hat{\delta}_1\mu_{\mathrm{educ}}$$

  where the second equality is the partial effect of female at the mean value of education from the regular model.

- Thus, the coefficient on female in the reparametrized model, $\hat{\alpha}_1$, is the wage gap at the mean value of education, $\mathrm{educ} = \mu_{\mathrm{educ}}$. We can use associated t statistic to decide whether it is significant or not.

- In R:

```
reg = lm(log(hwage) ~  female*educ + emp_sect, data=hls)
round(coef(summary(reg)),3)
```

Interactions IV

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.585      0.065  24.204    0.000
## female           -0.052      0.103  -0.507    0.612
## educ              0.059      0.005  12.361    0.000
## emp_sectother    -0.840      0.164  -5.137    0.000
## emp_sectpriv     -0.749      0.043 -17.230    0.000
## female:educ      -0.004      0.009  -0.436    0.663
```

- If we simply look at the coefficient on $female$, we may conclude wrongly that it is not significant at the 5% level. But this coefficient supposedly measures the effect when $educ = 0$, which is not interesting.

- To compute the partial effect of $female$ on wage at the mean value of $educ$, we can use directly the formula $\hat{\alpha}_1 = \hat{\delta}_0 + \hat{\delta}_1 \mu_{educ}$

```
mu_edu = mean(hls$educ)
mu_edu
```

```
## [1] 9.26
```

```
b <- coef(reg)
b
```

Interactions V

```
##   (Intercept)        female          educ emp_sectother  emp_sectpriv 
##       1.58529      -0.05216       0.05944      -0.84022      -0.74933 
##   female:educ 
##      -0.00393
```

```
b[["female"]] + mu_edu*b[["female:educ"]]
```

```
## [1] -0.0885
```

- But this does not give us the standard error of this new estimate. We need to rerun the regression, where we replace the interaction term with ”demeaned” interaction term.

```
reg2 = lm(log(hwage) ~ female+educ + female:I(educ-mu_edu)
           + emp_sect, data=hls)
round(coef(summary(reg2)),3)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.585      0.065  24.204    0.000
## female                     -0.089      0.043  -2.053    0.040
## educ                        0.059      0.005  12.361    0.000
## emp_sectother              -0.840      0.164  -5.137    0.000
## emp_sectpriv               -0.749      0.043 -17.230    0.000
## female:I(educ - mu_edu)    -0.004      0.009  -0.436    0.663
```

Interactions VI

- We see that the associated t statistic with $\hat{\alpha}_1 = -0.089$ is -2.053 which means that it is significant at th 5 % significance level.

Migrant wage gap I

- Assume that we have the following data generating process on migrant and native population

  $$wage^n = 300 + 10edu + 20age + u$$

  $$wage^m = 200 + 20edu + 15age + v$$

  Of course age and education distribution may differ across migrants and natives. More on that later.

  ```
  url = "https://github.com/obakis/econ_data/raw/master/fake_mig_dat.rds"
  download.file(url, "fake_mig_dat.rds", mode ="wb")
  mdat = readRDS("fake_mig_dat.rds")
  head(mdat)
  ```

- We want to know whether there exists a wage differential between migrants and natives

  ```
  mreg = lm(wage~migrant,data=mdat)
  summary(mreg)$coef
  ```

Migrant wage gap II

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     1355       9.11   148.7 3.12e-117
## migrant         -602      20.37   -29.6  1.25e-50
```

- The answer is Yes: native wages are higher by 602 on average, but age and education distributions are not the same across natives and migrants ...

```
aggregate(cbind(educ,age) ~ migrant, FUN=mean, data=mdat)
```

```
##   migrant  educ  age
## 1       0 15.05 45.2
## 2       1  5.25 29.9
```

```
summary(lm(age ~ migrant, data=mdat))$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     45.2       0.52    87.1 1.27e-94
## migrant        -15.4       1.16   -13.2 1.47e-23
```

```
summary(lm(educ ~ migrant, data=mdat))$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     15.1      0.247    60.9 1.16e-79
## migrant         -9.8      0.553   -17.7 2.52e-32
```

Migrant wage gap III

- and they matter for wage:

```
reg0 = lm(wage ~ age+educ , data=mdat)
summary(reg0)$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     12.7     22.213   0.573 5.68e-01
## age             21.5      0.687  31.219 2.13e-52
## educ            24.2      1.176  20.597 3.42e-37
```

- If we want to compute the wage differential between natives and migrants we need to control for educ and age differences. Given our data generating process that we used above, actually we can easily compute the expected wage difference between a native and a migrant who have the same (for instance average) age and education level.

- For two individuals with average education and age the true wage difference should be 180 approximately.

```
ave_age = mean(mdat$age) ; ave_age
```

```
## [1] 42.2
```

Migrant wage gap IV

```
ave_edu = mean(mdat$educ) ; ave_edu
```

```
## [1] 13.1
```

```
wn = 300 + 10*ave_edu + 20*ave_age # true wage for natives
wm = 200 + 20*ave_edu + 15*ave_age  # true wage for migrant
wn - wm
```

```
## [1] 180
```

- However, for observed data we do not know the data generating process. To control for age and education we assume a linear model. A basic one is the following:

```
reg1 = lm(wage~age+educ+migrant,data=mdat)
summary(reg1)$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    414.4     19.327    21.4 2.21e-38
## age             16.8      0.334    50.4 6.98e-71
## educ            11.9      0.701    17.0 1.27e-30
## migrant       -226.8      9.754   -23.3 3.19e-41
```

where the wage difference is only 227.

Migrant wage gap V

- This is far better than 602 (unconditional/unadjusted wage difference). However, the above model assumes that age and education have the same effect in native and migrant populations. This is not true.
- As a fits step let us assume that the return to education differ for migrants and natives.

```
reg2 = lm(wage ~ age + educ*migrant, data=mdat)
summary(reg2)$coef
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    431.16     14.949   28.84 8.78e-49
## age             17.15      0.259   66.29 2.40e-81
## educ             9.81      0.594   16.53 1.08e-29
## migrant       -302.53     11.816  -25.60 2.01e-44
## educ:migrant    11.48      1.387    8.27 7.93e-13
```

```
aw_m = predict(reg2, list(migrant=1, age=ave_age, educ=ave_edu))
aw_n = predict(reg2, list(migrant=0, age=ave_age, educ=ave_edu))
aw_n - aw_m
```

```
##   1
## 152
```

Migrant wage gap VI

```
##### To test for significance when educ = ave_edu
reg3 = lm(wage ~  age + educ + migrant +
             migrant:I(educ-ave_edu), data=mdat)
coef(summary(reg3))
```

```
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   431.16     14.949   28.84 8.78e-49
## age                            17.15      0.259   66.29 2.40e-81
## educ                            9.81      0.594   16.53 1.08e-29
## migrant                      -152.25     11.707  -13.01 8.16e-23
## migrant:I(educ - ave_edu)      11.48      1.387    8.27 7.93e-13
```

- A more general and better option is to allow the return to education and to age differ across migrants status. This is the same thing as running separate regressions for natives and migrants (except common intercept). Since the version with interaction terms can provide further information regarding migration effect we prefer the version with interaction terms

```
reg4n = lm(wage ~ age + educ, data=subset(mdat, migrant==0))
coef(summary(reg4n))
```

Migrant wage gap VII

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   308.56      8.457    36.5 2.37e-50
## age            19.82      0.165   120.4 1.99e-89
## educ            9.95      0.247    40.4 1.47e-53
```

```
reg4m = lm(wage ~ age + educ, data=subset(mdat, migrant==1))
coef(summary(reg4m))
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    196.6      5.856    33.6 5.57e-17
## age             15.2      0.146   103.7 2.92e-25
## educ            19.6      0.545    36.0 1.75e-17
```

```
aw_n = predict(reg4n, list(age=ave_age, educ=ave_edu))
aw_m = predict(reg4m, list(age=ave_age, educ=ave_edu))
aw_n-aw_m # we can not know whether the diff is significant or not
```

```
##   1
## 181
```

```
reg5 = lm(wage ~ (age + educ)*migrant, data=mdat)
summary(reg5)$coef
```

Migrant wage gap VIII

```
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    308.56      8.506    36.3  4.49e-57
## age             19.82      0.166   119.7 1.53e-104
## educ             9.95      0.248    40.1  6.05e-61
## migrant       -111.98     10.244   -10.9  1.96e-18
## age:migrant     -4.64      0.219   -21.2  1.25e-37
## educ:migrant     9.64      0.586    16.4  2.10e-29
```

```
aw_m = predict(reg5, list(migrant=1, age=ave_age, educ=ave_edu))
aw_n = predict(reg5, list(migrant=0, age=ave_age, educ=ave_edu))
aw_n - aw_m # we can not know whether the diff is significant or not
```

```
##   1
## 181
```

```
##### To test for significance
reg6 = lm(wage ~  age + educ + migrant +
             migrant:I(age-ave_age) +
             migrant:I(educ-ave_edu), data=mdat)
coef(summary(reg6))
```

Migrant wage gap IX

```
##                           Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                 308.56      8.506    36.3  4.49e-57
## age                          19.82      0.166   119.7 1.53e-104
## educ                          9.95      0.248    40.1  6.05e-61
## migrant                    -181.46      5.080   -35.7  1.74e-56
## migrant:I(age - ave_age)     -4.64      0.219   -21.2  1.25e-37
## migrant:I(educ - ave_edu)     9.64      0.586    16.4  2.10e-29
```

The more flexible model gives a wage difference equal to 181.5 which is really close to 180.

Outline

1. Regression analysis: introduction

2. Regression analysis: further issues
   - Inference: theoretical background
   - Hypothesis testing about a single parameter
   - Testing multiple linear restrictions: the F test

Normality assumption I

- Up to here, we showed that the OLS estimators are BLUE. This is useful in describing the precision (mean and variance) of the OLS estimators but not sufficient for statistical inference which requires the full sampling distribution.

- Gauss-Markov assumptions do not imply a distribution for the OLS estimators. The shape of the distribution of the OLS estimators, is determined by the distribution of the error term, u.

- **Assumption 6**: We assume that the unobserved error is normally distributed

- The error is independent of independent variables and is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0, \sigma^2)$

Normality assumption II

- Assumption 6 (normality) is much stronger than our previous assumptions. It implies also assumption 4 (zero conditional mean) and assumption 5 (homoskedasticity).

- For cross-sectional data, Assumptions 1-6 are called the classical linear model (CLM) assumptions.

- Gauss-Markov assumptions $+$ normality $=$ CLM

- How is it justifiable to assume normality for u?

  - **central limit theorem:** the arithmetic mean of a sufficiently large number of independent random variables, each with a well-defined (finite) expected value and finite variance, will be approximately normally distributed, regardless of the underlying distribution.

  - Since u is the sum of many factors affecting y, one may conclude that u should has an approximate normal distribution.

Normality assumption III

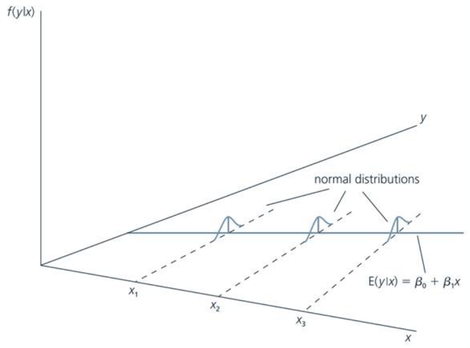
  - The central limit theorem suggests that each component of u affect y in an additive fashion. This may not be true in practice.
  - Normality of u is an empirical matter: using $wage$ vs $\log(wage)$ and other transformations where necessary.
  - When we have large samples, nonnormality of the errors may not be a serious problem (asymptotic normality !!!).

- Normality is about the distribution of u not y or x. In general neither y, nor x would be normally distributed in practice. But it is OK to say that conditional on x, y has a normal distribution.

Normality assumption IV

FIGURE 4.1

The homoskedastic normal distribution with a single explanatory variable.



Normality assumption V

Theorem (Normal sampling distributons):

$$\hat{\beta}_j \sim \text{Normal}[\beta_j, \text{Var}(\hat{\beta}_j)] \quad \Rightarrow \frac{\hat{\beta}_j - \beta_j}{\text{sd}(\hat{\beta}_j)} \sim \text{Normal}(0, 1)$$

- The standardized estimator makes use of sd which relies on the unknown constant σ. When we use the estimate $\hat{\sigma}$, we get the t distribution with $n - (k + 1)$ degrees of freedom (df) instead of standard normal distribution (SND)

  $$\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-k-1}$$

  A t distribution is similar to a SND, it is heavier in the tails, flatter near the center, and its exact dispersion is dictated by df. When number of observation is large ($n \geqslant 40$) the t distribution is well approximated by the SND.

Normality assumption VI

- What does it mean ”normally distributed” with a given mean and sd? Ex. What is the probability that a number we draw from a random process with $\mathrm{mean} = 70$ and $\mathrm{sd} = 4$ is between 62 and 78 (within 2 standard deviations of the mean) ?

```r
set.seed(1279)
x=rnorm(n=10, mean=70,sd=4)
round(x,1)
```

```
##  [1] 70.1 68.3 65.2 67.0 67.0 70.8 66.8 69.2 59.5 67.2
```

```r
sum(x >= 62 & x <= 78)/10
```

```
## [1] 0.9
```

```r
x2=rnorm(n=20, mean=70,sd=4)
#round(x2,1)
sum(x2 >= 62 & x2 <= 78)/20
```

```
## [1] 1
```

```r
x3=rnorm(n=50, mean=70,sd=4)
sum(x3 >= 62 & x3 <= 78)/50
```

```
## [1] 0.94
```

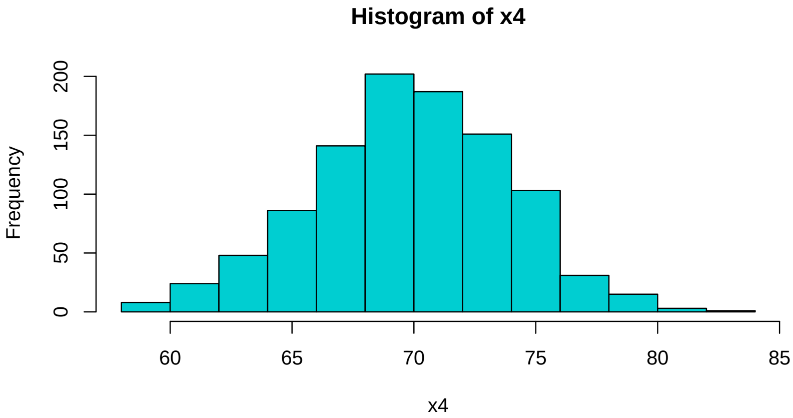
Normality assumption VII

```r
x4=rnorm(n=1000, mean=70,sd=4)
sum(x4 >= 62 & x4 <= 78)/1000
```

```
## [1] 0.949
```

```r
hist(x4,col="darkturquoise")
```

Normality assumption VIII



- Standardization is just a rescaling:

Normality assumption IX

```
val = c(62,70,78) # mean=70,sd=4
vals = (val-70)/4 # mean=0,sd=1
vals # How many std deviation each number is far from the mean?
```

```
## [1] -2  0  2
```

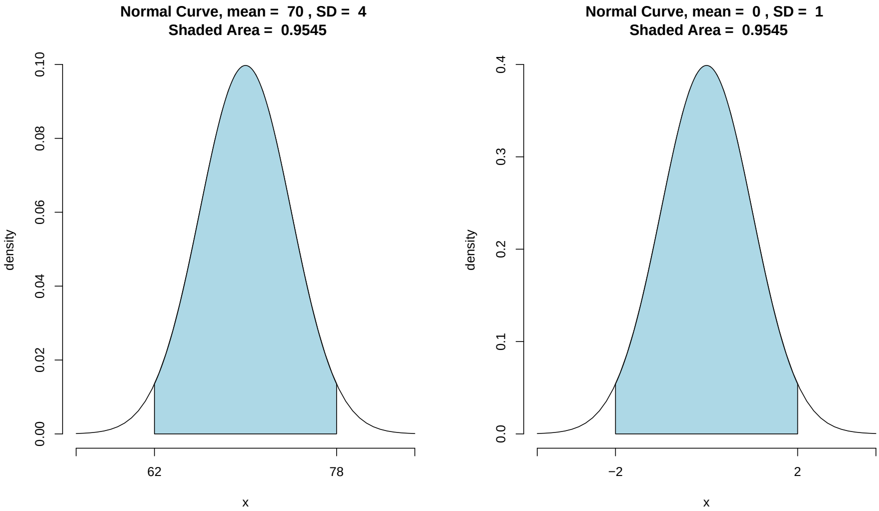
- The probability that a number we draw a number from a normal distribution so that this number is within 2 standard deviations of the mean is approximately 95%. This is always true, whether the random variable is standardized or not does not matter.

- For x these values are $mean = 70$ and $sd = 4$. So, the boundary values are $62 = 70 - 2 \times 4$ and $78 = 70 + 2 \times 4$. For standardized x these values are $mean = 0$ and $sd = 1$. And the boundary values are $-2 = 0 - 2 \times 1$ and $2 = 0 + 2 \times 1$.

$$P(62 \leqslant x \leqslant 78) = P\left(-2 \leqslant \frac{x - \bar{x}}{sd(x)} \leqslant 2\right) \approx 0.95$$

If we use a continuous density function we get the following:

Normality assumption X



t test I

- It is important to remember that we do not observe β, we compute $\widehat{\beta}$ under Gauss-Markov assumptions and we would like to infer the value of β using our estimate $\widehat{\beta}$. Since knowing the exact value of β would be difficult we try to assess, at least, the likelihood of particular values such as $H_0 : \beta_j = a$.

- Imagine for a moment we know β_j (the unknown population parameter). Even if we know β_j it is rarely the case that we get exactly $\widehat{\beta}_j = \beta_j$ because of **sampling distribution**. We will have a distribution for $\widehat{\beta}_j$: $\widehat{\beta}_j \sim \text{Normal}[\beta_j, \text{Var}(\widehat{\beta}_j)]$

t test II

- And usually, in practice, $\hat{\beta}_j \neq \beta_j$. But we know that approximately 95% of the time $\hat{\beta}_j$ will be between $\beta_j - 2 \times se(\hat{\beta}_j)$ and $\beta_j + 2 \times se(\hat{\beta}_j)$, or equivalently, $\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)}$ is between -2 and 2

$$P\left(-2 \leqslant \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \leqslant 2\right) \approx 0.95$$

- The problem is that we do not know β_j !!! One idea is to assume a given value for it, $H_0 : \beta_j = a$, and then decide whether the observed $\hat{\beta}_j$ is consistent with $H_0 : \beta_j = a$ or not. Thus, we are inferring the value of unobserved parameter β_j from the observed estimate $\hat{\beta}_j$.

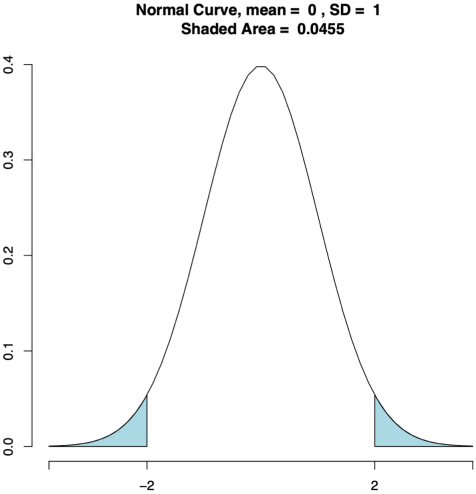
t test III

- This is where t **statistic** intevenes. It measures how many standard errors $\hat{\beta}_j$ is away from a, the ”hypothesized value” of the unknown parameter.

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - a}{\text{se}(\hat{\beta}_j)}$$

- If $H_0 : \beta_j = a$ is true, then the **expected value** for the t statistic would be zero and 95% of the time the t statistic will be between -2 and 2. This means that we are confident in H_0.
- If H_0 is wrong ($\beta_j \neq a$) then t statistic will move away from zero. The farther the t statistic is from the zero the less confident we are that H_0 is likely to be true. **The question is how far?**

t test IV



t test V

- We need such a rule: if observed t statistic is higher than c in absolute value, $|t| > c$, we reject H_0. In practice, one frequently used value for c is 2 which corresponds to a 5% significance level

- Actually by deciding on a critical value, we also decide on **significance level**. Even if we reject H_0 (because $|t| > c$), it is still possible that H_0 is true. This probability is equal to the significance level we use.

- Thus, the significance level is the probability of rejecting $H_0 : \beta_j = a$ when it is in fact true. $\Rightarrow$ The risk we are taking in rejecting H_0 when it is in fact true!

- **Remark:** By default, $a = 0$ and the reported t statistic is equal to $\frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}$, in all(?) modern econometrics software programs.

t test VI

- **Example:** Assume that there is a normally distributed random series with $mean = 4$ and $sd = 1.5$

  ```r
  set.seed(1270)
  x=rnorm(n=1000, mean=4, sd=1.5)
  ```

- We assume that the true mean is unknown to us (econometrician). We want to infer the true mean through the usual t statistic. We try two different values for α: $H_0 :\ \text{mean} = 3.5$ and $H_0 :\ \text{mean} = 0$.

  ```r
  t1 = (x-3.5)/1.5
  mean(t1) ## mean of t1, should be close to 0 if null is true
  ```

  ```
  ## [1] 0.398
  ```

  ```r
  sd(t1)
  ```

  ```
  ## [1] 0.974
  ```

  ```r
  t2 = (x-0)/1.5
  mean(t2) ## mean of t2, should be close to 0 if null is true
  ```

  ```
  ## [1] 2.73
  ```

  ```r
  sd(t2)
  ```

t test VII

```
## [1] 0.974
```

- **Remark:** In reality, we do not have as many t statistics as above. We have only an estimate and a t statistic for it. We can think that we draw randomly 1 t statistic from both t_1 and t_2, this would be our ”observed t statistic”.

- When $H_0 : \text{ mean} = 3.5$ (case of t_1) most realizations of t statistics are centered around zero, this is why we are likely to observe a t statistic between -2 and 2 (95%). Thus, we are confident in $H_0 : \text{ mean} = 3.5$.

```
sum(t1 >= -2.0 & t1 <= 2.0)/1000
```

```
## [1] 0.942
```

- But when $H_0 : \text{ mean} = 0$ (case of t_2) most realizations of t statistics are centered around 3, which should not happen if H_0 is true. So, we will most likely reject $H_0 : \text{ mean} = 0$ because only a small part of ”observed t statistic” values between -2 and 2 (22%).

```
sum(t2 >= -2.0 & t2 <= 2.0)/1000
```

t test VIII

```
## [1] 0.223
```

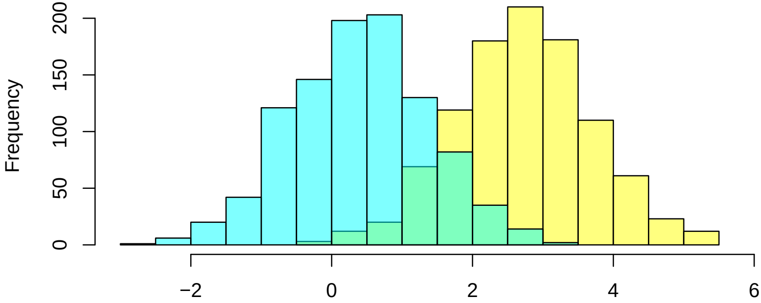
```
sum( t2 >= 2.0)/1000
```

```
## [1] 0.777
```

- As a result, we reject easily $H_0: \text{ mean} = 0$ while $H_0: \text{ mean} = 3.5$ can not be rejected.
- Graphs for t_1 and t_2

```
hist(t2, col=rgb(1,1,0,0.5), xlim=c(-3,6), xlab="", main = "")
hist(t1, col=rgb(0,1,1,0.5), add=TRUE)
```

t test IX



t test X

- In practice, we use 10% or 5% or 1% as **significance level** (α). Most studies prefer 5% is which means that choose to mistakenly reject H_0 when it is true 5% of the time.

- 2 factors that will determine the precise rejection rule are the alternative hypothesis and the chosen **significance level** of the test.

- The widely used alternative hypothesis is $H_1 \neq 0$ (2-sided alternative), but $H_1 > 0$ and $H_1 < 0$ are used as well (1-sided alternative).

- Say, our alternative hypothesis is two-sided, $H_1 \neq 0$ and we choose 5 % as significance level (α). This α and H_1 imply a critical value c, given df. Then we reject H_0 only if $|t_{\hat{\beta}_j}| > c$.

- The values of c are traditionally published in statistical tables.

t test XI

- In modern times, all we need is a call to `qt()` function in R.
- We want to know for which positive number c the area under the t **distribution** between $-c$ and c is 0.95, $P(-c \leqslant X \leqslant c) = 0.95$.
    - If 95% of the area lies between $-c$ and c, then 5% of the area must lie outside of this range. Given that the t distribution is symmetric, half of this amount, 2.5%, must lie before $-c$. Then the area under the curve before c must be: $0.025 + 0.95 = 0.975$, thus $P(X \leqslant c) = 0.975$.
    - The boundary value that gives an area of 97.5%, under the t distribution, is 1.98 (we assume $df = 120$)

```
#qt(0.975, df=120) #qnorm(0.975,mean=0,sd=1)
alfa = 5/100
qt(1- alfa/2, df=120) #qnorm(0.975,mean=0,sd=1)
```

```
## [1] 1.98
```

Two-sided alternative: example I

Econometric model: $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$

```r
url = "https://github.com/obakis/econ_data/raw/master/wage1.rds"
download.file(url, "wage1.rds", mode ="wb")
wage1 = readRDS("wage1.rds")
View(wage1)


reg1 = lm(lwage ~ educ + exper + tenure, data=wage1)
summary(reg1)
```

```
##
## Call:
## lm(formula = lwage ~ educ + exper + tenure, data = wage1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.0580 -0.2965 -0.0326  0.2879  1.4281
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

Two-sided alternative: example II

```
## (Intercept)  0.28436    0.10419    2.73   0.0066 **
## educ         0.09203    0.00733   12.56  < 2e-16 ***
## exper        0.00412    0.00172    2.39   0.0171 *
## tenure       0.02207    0.00309    7.13  3.3e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.441 on 522 degrees of freedom
## Multiple R-squared:  0.316,Adjusted R-squared:  0.312
## F-statistic: 80.4 on 3 and 522 DF,  p-value: <2e-16
```

- $n - (k + 1) = 526 - (3 + 1) = 522$;

  ```
  df_ex1 = df.residual(reg1)
  df_ex1
  ```

  ```
  ## [1] 522
  ```

- 5% significance level: $c = 1.96$;

Two-sided alternative: example III

```
alfa = 0.05
# alfa/ 2 for 2-sided, alfa for 1-sides
c_ex1=qt(1-alfa/2, df=df_ex1)
c_ex1
```

```
## [1] 1.96
```

- $t_{exper} = \frac{0.0041}{0.0017} \cong 2.4 > 1.96$, etc.

```
t_ex1 = summary(reg1)$coef[,3]
t_ex1 > c_ex1
```

```
## (Intercept)        educ       exper      tenure
##        TRUE        TRUE        TRUE        TRUE
```

- Thus, $H_0 : \beta_j = 0$ is rejected at the 5% significance level for all covariates and constant.
- Interpretation: $exper$ (or $\hat{\beta}_{exper}$) is **statistically significant at the the 5% significance level**

Two-sided alternative: example IV

- **Remark:** $n - (k+1) \geqslant 120$, which implies that we can use Normal distribution instead of t distribution.

Some remarks on testing I

1. Either we reject H_0 or fail to reject it. We never ”accept” H_0. When we cannot reject $H_0 : \beta = 0$, it is unlikely that we can reject $H_0 : \beta = \varepsilon$ where ε is an arbitrary small number. Since all of these non-rejected values of ε cannot be true at the same time, we say ” we fail to reject H_0” instead of ”we accept H_0”.

2. Only **UNDER** H_0, the t statistic has a t distribution with $n - (k + 1)$ df. We never know what the true distribution of $\hat{\beta}$ under H_1 !!!

3. When we have large number of observations (in practice this means $\geqslant 40$) t distribution converges to a std. normal distr. whose 97.5^{th} percentile is very close to 2 so that we use the following simple rule of thumb $|t_{\hat{\beta}_j}| \geqslant 2$ for t tests.

4. $H_0 : \hat{\beta}_j = 0$ is not meaningful. Why?

5. If we cannot reject H_0, we say: x_j has no partial (ceteris paribus) effect on y.

Confidance intervals I

- If we repeat the process and get a fairly large number of estimates for β_j we could construct a confidence interval (CI) for the unknown population parameter β_j directly. But we have only one sample !

- Once we compute $\hat{\sigma}$ (an estimate for error variance), we know that $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j)$ has a t distribution with $n-(k+1)$ df. This means that 95% of the time $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j)$ will be between $-c$ and c (for large samples $c \approx 2$). Using this we get 95% CI for β_j as:

$$0.95 = P\left(-c \leqslant \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \leqslant c\right) = P(-c \times se(\hat{\beta}_j) \leqslant \hat{\beta}_j - \beta_j \leqslant c \times se(\hat{\beta}_j))$$

$$= P(\hat{\beta}_j - c \times se(\hat{\beta}_j) \leqslant \beta_j \leqslant \hat{\beta}_j + c \times se(\hat{\beta}_j))$$

  where c is determined by the specified significance level.

- If we want a CI of 95%, we need to find c that corresponds to the 97.5^{th} percentile in a t distribution with $n-(k+1)$ df.

Confidance intervals II

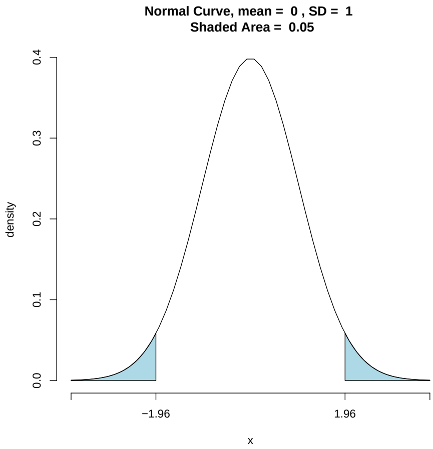
- For the standard normal distribution the 97.5^{th} percentile is 1.96 which is very close to 2. For t distribution it depends on degrees of freedom, $df = n - k - 1$, but when $df \geqslant 40$ the t distribution is close to he standard normal distribution so that for a CI of 95% we use the following simple rule of thumb

$$\underline{\beta}_j \equiv \hat{\beta} - 2 \times se(\hat{\beta})$$

$$\overline{\beta}_j \equiv \hat{\beta} + 2 \times se(\hat{\beta})$$

to find the lower and upper bounds of the CI in practice.

Confidance intervals III



the 97.5^{th} percentile in a normal distribution is 1.96.

Confidance intervals IV

- How to interpret CI? Think as follows: we draw 100 random samples from population and estimate $\hat{\beta}_j$ 100 times. 95 times our estimate $(\hat{\beta})$ for unknown parameter (β) will reside in the $(\underline{\beta}_j, \overline{\beta}_j)$ interval.

- Of course, the above is the interpretation. In reality we compute CI from unique sample we have using our estimate and residuals. Can we be sure that this unique sample is part of the above 95%? Unfortunately NO!

- Once we have our CI, we can test for $H_0 : \beta_j = a$ in favor of $H_1 : \beta_j \neq a$. If a lies in the CI we fail to reject H_0, but reject it whenever a is not in the CI.

- Consider the following model

$$\log RD = \beta_0 + \beta_1 \log(sales) + \beta_2 profmarg + u$$

  where we are explaining R&D expenditures by sales and profit margin (ratio of profits to sales)

Confidance intervals V

- Can we reject $H_0 : \beta_0 = 0$ or $H_0 : \beta_1 = 1$ or $H_0 : \beta_2 = 0$?

```r
url = "https://github.com/obakis/econ_data/raw/master/rdchem.rds"
download.file(url, "rdchem.rds", mode ="wb")
rdchem = readRDS("rdchem.rds")
View(rdchem)
```

```r
reg <- lm(log(rd) ~ log(sales)+profmarg, data=rdchem)
summary(reg)
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.3783     0.4680   -9.35  2.9e-10 ***
## log(sales)    1.0842     0.0602   18.01  < 2e-16 ***
## profmarg      0.0217     0.0128    1.69      0.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.514 on 29 degrees of freedom
## Multiple R-squared:  0.918,Adjusted R-squared:  0.912
## F-statistic:  162 on 2 and 29 DF,  p-value: <2e-16
```

Confidance intervals VI

- With $n - (k+1) = 32 - 3 = 29$ df we have $c = 2.045$ at the 5% significance level (95% confidence level). The CI for $\log(sales)$ is computed as $(\hat{\beta}_1 = 1.084 \mp 2.045 \times 0.06) \equiv (0.961, 1.21)$.

- Since zero does not lie in the CI for $\log(sales)$ we can safely reject $H_0 : \beta_1 = 0$ at 5% level.

- Please remark that we can not reject, for instance, $H_0 : \beta_1 = 1$ or $H_0 : \beta_1 = 1.1$, or $H_0 : \beta_1 = 0.98$, at 5% level for the variable $\log(sales)$.

- Following the same steps, we can show that the CI for variable $profmarg$ is given by $(-0.0045, 0.0479)$. Given that zero is included in the 95% confidence interval, we fail to reject $H_0 : \beta_2 = 0$ at the 5% level.

Confidance intervals VII

```
confint(reg) # by default, CI=0.95
```

```
##                 2.5 %  97.5 %
## (Intercept) -5.33548 -3.4211
## log(sales)   0.96111  1.2073
## profmarg    -0.00449  0.0478
```

```
confint(reg, level=0.99)
```

```
##                0.5 %  99.5 %
## (Intercept) -5.6683 -3.0882
## log(sales)   0.9183  1.2501
## profmarg    -0.0136  0.0569
```

F test: exclusion restrictions I

- Consider the following unrestricted model

  $$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

  where $salary$ is the 1993 total salary, $years$ is years in the league, $gamesyr$ is average games played per year, $bavg$ is career batting average, $hrunsyr$ is home runs per year, and $rbisyr$ is runs batted in per year.

- We want to test $H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$. The null hypothesis we test is that, once years in the league and games per year have been controlled for, the statistics measuring performance (bavg, hrunsyr, and rbisyr) have no effect on salary.

- These are called **exclusion restrictions**. There are 3 such restrictions in the null ($q = 3$). Imposing them, we get the restricted model.

F test: exclusion restrictions II

- The above exclusion restrictions are an example of a set of **multiple restrictions** because we are putting more than one restriction on parameters.

- A test of multiple restrictions is called a **multiple hypotheses test** or a **joint hypotheses test**.

- F statistic is given by

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(R^2_{ur} - R^2_r)/q}{(1-R^2_{ur})/(n-k-1)}$$

- Since R-squared from unrestricted will be at least as high as high as R-squared from restricted regression, F statistic is always positive !

F test: exclusion restrictions III

- F statistic follows an F distribution with q degrees of freedom in the numerator and $n-(k+1)$ degrees of freedom in the denominator, under the null hypothesis, and assuming that the CLM assumptions hold.

```r
url = "https://github.com/obakis/econ_data/raw/master/mlb1.rds"
download.file(url, "mlb1.rds", mode ="wb")
mlb1 = readRDS("mlb1.rds")
View(mlb1)

# Unrestricted OLS regression:
reg_ur <- lm(log(salary) ~ years+gamesyr+bavg+hrunsyr+rbisyr, data=mlb1)
summary(reg_ur)
```

F test: exclusion restrictions IV

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.12e+01   2.89e-01   38.75  < 2e-16 ***
## years       6.89e-02   1.21e-02    5.68  2.8e-08 ***
## gamesyr     1.26e-02   2.65e-03    4.74  3.1e-06 ***
## bavg        9.79e-04   1.10e-03    0.89     0.38
## hrunsyr     1.44e-02   1.61e-02    0.90     0.37
## rbisyr      1.08e-02   7.17e-03    1.50     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.727 on 347 degrees of freedom
## Multiple R-squared:  0.628,Adjusted R-squared:  0.622
## F-statistic:  117 on 5 and 347 DF,  p-value: <2e-16
```

Using ”car” package, F testing is as simple as:

```
library(car)
linearHypothesis(reg_ur, c("bavg=0","hrunsyr=0","rbisyr=0"))
```

F test: exclusion restrictions V

```
## Linear hypothesis test
##
## Hypothesis:
## bavg = 0
## hrunsyr = 0
## rbisyr = 0
##
## Model 1: restricted model
## Model 2: log(salary) ~ years + gamesyr + bavg + hrunsyr + rbisyr
##
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1    350 198
## 2    347 183  3      15.1 9.55 4.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F test: general linear restrictions I

It is also possible to test for general linear restrictions apart from exclusion restrictions.

- Consider the following example

  $\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \beta_4\, bdrms + u$

  where $price$ is house price, $assess$ is the assessed housing value (before the house was sold), $lotsize$ is size of the lot, in feet, $sqrft$ is square footage and finally $bdrms$ number of bedrooms.

- Suppose we would like to test whether the assessed housing price is a rational valuation. If this is the case, then a 1% change in $assess$ should be associated with a 1% change in $price$ and all other factors should be not related with house price, once the assessed value has been controlled for. The rationality assumption would be

$$H_0 : \beta_1 = 1, \quad \beta_2 = 0, \quad \beta_3 = 0, \quad \beta_4 = 0$$

F test: general linear restrictions II

- The last 3 are exclusion restrictions while $\beta_1 = 1$ is a linear restriction different from exclusion restrictions.

- Using the car package

```r
url = "https://github.com/obakis/econ_data/raw/master/hprice1.rds"
download.file(url, "hprice1.rds", mode ="wb")
hprice1 = readRDS("hprice1.rds")
View(hprice1)

# Unrestricted OLS regression:
reg_ur <- lm(lprice ~ lassess+llotsize+lsqrft+bdrms, data=hprice1)
summary(reg_ur)
```

F test: general linear restrictions III

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.26374    0.56966    0.46     0.64
## lassess      1.04307    0.15145    6.89    1e-09 ***
## llotsize     0.00744    0.03856    0.19     0.85
## lsqrft      -0.10324    0.13843   -0.75     0.46
## bdrms        0.03384    0.02210    1.53     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.148 on 83 degrees of freedom
## Multiple R-squared:  0.773,Adjusted R-squared:  0.762
## F-statistic: 70.6 on 4 and 83 DF,  p-value: <2e-16
```

```
library(car)
linearHypothesis(reg_ur, c("lassess=1", "llotsize=0", "lsqrft=0", "bdrms=0"))
```

F test: general linear restrictions IV

```
## Linear hypothesis test
##
## Hypothesis:
## lassess = 1
## llotsize = 0
## lsqrft = 0
## bdrms = 0
##
## Model 1: restricted model
## Model 2: lprice ~ lassess + llotsize + lsqrft + bdrms
##
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     87 1.88
## 2     83 1.82  4    0.0586 0.67   0.62
```

So we fail to reject H_0.

Heteroskedasticity robust std. erros I

```
library(lmtest) # for coeftest
coeftest(reg1) # assuming homoskedasticity

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.28436    0.10419    2.73   0.0066 **
## educ         0.09203    0.00733   12.56  < 2e-16 ***
## exper        0.00412    0.00172    2.39   0.0171 *
## tenure       0.02207    0.00309    7.13  3.3e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

library(sandwich) # for vcovHC
coeftest(reg1, vcov = vcovHC) # heteroskedasticity robust, R default: "HC3"
```

Heteroskedasticity robust std. erros II

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.28436    0.11331    2.51    0.012 *
## educ         0.09203    0.00804   11.44  < 2e-16 ***
## exper        0.00412    0.00176    2.34    0.020 *
## tenure       0.02207    0.00386    5.72  1.8e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Heteroskedasticity robust std. erros III

```
coeftest(reg1, vcov = vcovHC(reg1, "HC3")) # robust, R default: "HC3"
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.28436    0.11331    2.51    0.012 *
## educ         0.09203    0.00804   11.44  < 2e-16 ***
## exper        0.00412    0.00176    2.34    0.020 *
## tenure       0.02207    0.00386    5.72  1.8e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Heteroskedasticity robust std. erros IV

```
coeftest(reg1, vcov = vcovHC(reg1, "HC1")) # robust, Stata default: "HC1"
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.28436    0.11171    2.55    0.011 *  
## educ         0.09203    0.00792   11.62  < 2e-16 ***
## exper        0.00412    0.00175    2.36    0.019 *  
## tenure       0.02207    0.00378    5.83  9.5e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```