The Inference and Analysis of (Linear) Regression

Babak Rezaee Daryakenari
The Institute of Political Science
Leiden University
s.rezaeedaryakenari@fsw.leidenuniv.nl
For more research methods handouts:
https://babakrezaee.github.io

Spring 2020

OLS regression: an example

Remember Acemoglu et al. (2001) that we have been using in previous session.

```
myData=read.csv("https://raw.githubusercontent.com/babakrezaee/MethodsCourses/master/LeidenUniv_MAQM2020/D

#Drop the missing values
myData=na.omit(myData, cols=c("avexpr", "logpgp95"))

#rawData = data.frame(avexpr=sort(rawData$avexpr),logpgp95=rawData$logpgp95)
myData <- myData[order(myData$avexpr), ]

library(jtools)
# OLS model
OLS1_results <- lm(logpgp95~avexpr, data=myData)
OLS2_results <- lm(logpgp95~avexpr+cons1, data=myData)
```

- What is the difference between a 0.0 and a .22 p-value?
- What is *p-value*?
- How does it affect our inference of the estimated association between two variables?

Results of model 1

```
#print results
summ(OLS1_results)
```

```
## MODEL INFO:
## Observations: 57
## Dependent Variable: logpgp95
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,55) = 72.32, p = 0.00
## R² = 0.57
## Adj. R² = 0.56
##
## Standard errors: OLS
## -----------------------------------------------
##                      Est.   S.E.   t val.      p
## ----------------- ------ ------ -------- ------
## (Intercept)          4.89   0.38    12.79   0.00
## avexpr               0.49   0.06     8.50   0.00
## -----------------------------------------------
```

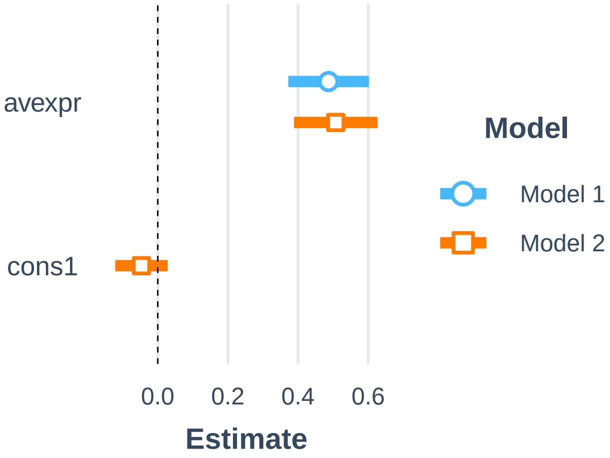
Results of model 2

```r
#print results
summ(OLS2_results)
```

```
## MODEL INFO:
## Observations: 57
## Dependent Variable: logpgp95
## Type: OLS linear regression
##
## MODEL FIT:
## F(2,54) = 37.32, p = 0.00
## R² = 0.58
## Adj. R² = 0.56
##
## Standard errors: OLS
## ------------------------------------------------
##                      Est.   S.E.   t val.      p
## ----------------- ------- ------ -------- ------
## (Intercept)          4.92   0.38    12.91   0.00
## avexpr               0.51   0.06     8.56   0.00
## cons1               -0.05   0.04    -1.25   0.22
## ------------------------------------------------
```

Plotting the regression results

```r
## You need "ggstance" package installed for plot_summs
plot_coefs(OLS1_results, OLS2_results,  scale = FALSE,
           inner_ci_level = .95)
```

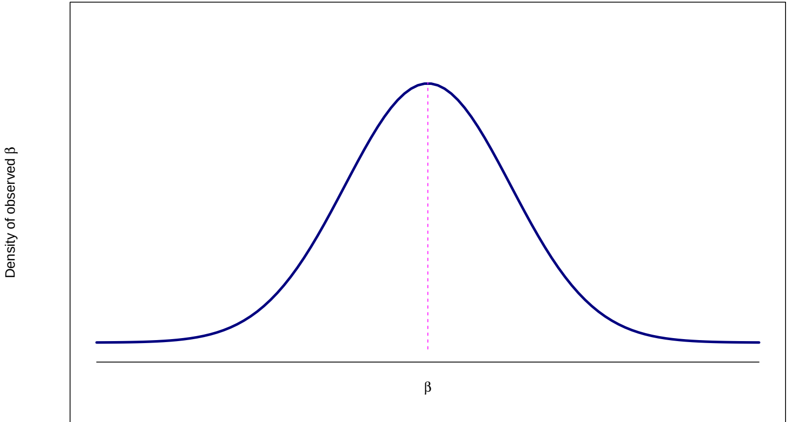


The distribution of $\hat{\beta}$

- We use an estimation of β, i.e. $\hat{\beta}$, because we do not know what is the *true* value of β.
- Therefore, we use a *sample* of data to evaluate an association between two variables.
- Working with a sample of the data, instead of the entire population of data, leads to an uncertainty.
- Statistical methods allow you to measure this uncertainty and decide how to interpret it: supporting your theory or rejecting your theory.

The distribution of $\hat{\beta}$ (2)

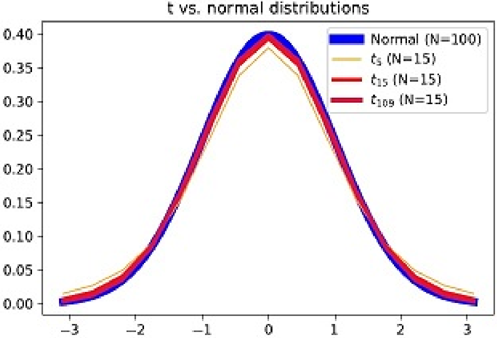
- If for different samples, we get different $\hat{\beta}$s, then we have a distribution of estimated $\hat{\beta}$s.
- Statistical models show that for large Ns, i.e. large samples, the distribution of $\hat{\beta}$ is normal: $\hat{\beta} \sim Normal(\beta, \sigma^2_{\beta})$



Thank you Guinness!

- We need a large sample to have a normal distribution and close to population variance.
- William Sealy Gosset in Guinness Brewery lab in Dublin developed a distribution that reaches to a normal distribution with much smaller sample. This distribution is known as t-distribution.
- Guinness researchers were not allowed to publish their research using their real name, so William Sealy Gosset signed his paper *t student*! And, that is why this distribution is called t-distribution!

t-distribution for different degrees of freedoms



Hypothesis testing

- $\hat{\beta}$ follows a *t*-distribution. How does this can help?
- We need go back to our first session. In a scientific study, we want to reject our theoretical hypothesis.
- Our theoretical hypothesis is that two variables are associated. That is, $\hat{\beta} \neq 0$.
- Since we want to attack to our hypothesis, we put $\hat{\beta} = 0$ as the null hypothesis, H_0. And, put our theoretical claim as the alternative hypothesis, H_a.

$$\begin{cases} H_0 : \hat{\beta} = 0 \\ H_a : \hat{\beta} \neq 0 \end{cases}$$

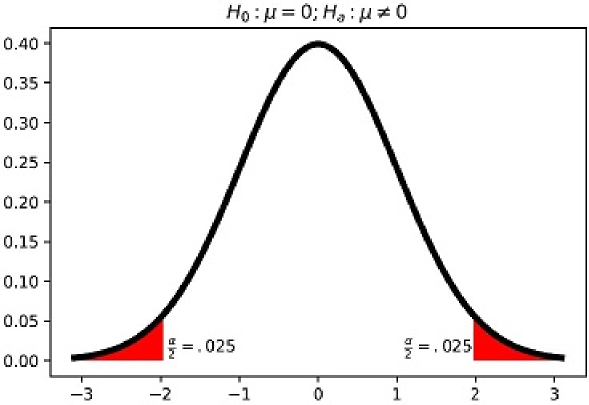
- Therefore, if we reject the null hypothesis of $\hat{\beta} = 0$, we find *support* for our hypothesis that $\hat{\beta} \neq 0$.
- That is why you should avoid writing that the estimations prove(!!!) your theoretical arguments.

The distribution of $H_0 : \hat{\beta} = 0$ and critical values

- We need a criteria to agree that the estimated $\hat{\beta}$ is *far enough* from zero!
- This is called critical value and shown by α.
- Common *alpha* values are 1%, 5%, and 10%, which receptively represents the famous $***$, $**$, and $*$ in regression tables.
- A critical value of 5% means that we have 95% confidence in our findings. If this analysis is repeated 100 times, 95 time we get similar results!

The distribution of $H_0 : \hat{\beta} = 0$ and critical values

- The distribution of $\hat{\beta} = 0$ with 5% critical value:



- If we estimate a model that its $\hat{\beta}$ falls in the critical value areas, i.e. $p - value < 5\%$, then we say the null hypothesis of $\beta = 0$ is rejected.

Getting back to our example

```
## MODEL INFO:
## Observations: 57
## Dependent Variable: logpgp95
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,55) = 72.32, p = 0.00
## R² = 0.57
## Adj. R² = 0.56
##
## Standard errors: OLS
## -----------------------------------------------
##                      Est.   S.E.   t val.      p
## ----------------- ------ ------ -------- ------
## (Intercept)          4.89   0.38    12.79   0.00
## avexpr               0.49   0.06     8.50   0.00
## -----------------------------------------------
```

```
## MODEL INFO:
## Observations: 57
## Dependent Variable: logpgp95
## Type: OLS linear regression
##
## MODEL FIT:
## F(2,54) = 37.32, p = 0.00
## R² = 0.58
## Adj. R² = 0.56
##
## Standard errors: OLS
## ------------------------------------------------
##                       Est.   S.E.   t val.      p
## ----------------- ------- ------ -------- ------
## (Intercept)           4.92   0.38    12.91   0.00
## avexpr                0.51   0.06     8.56   0.00
## cons1                -0.05   0.04    -1.25   0.22
## ------------------------------------------------
```

Plotting our results