Lab 03

## setup

setwd("C:/Users/22700/Desktop")  
library(data.table)  
library(ggplot2)  
library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

## Load the data

## You can use the list files command to check if the data file is stored in your working directory.

list.files()

## [1] "~$»ªºÆÄÜÔ´»·±£¼¯ÍÅ¹É·ÝÓÐÏÞ¹«Ë¾£¨871298£©¼ò±¨.docx"   
## [2] "~$°²£º2018ÄêÄê¶È±¨¸æ.docx"   
## [3] "~$¾©¿Ø.docx"   
## [4] "~$À¬»ø·¢µçÕþ²ß.docx"   
## [5] "~$ÊÐÉÌÒµ¼Æ»®Êé0319.docx"   
## [6] "~$ird Meeting Note.docx"   
## [7] "~$search Proposal Second Draft.docx"   
## [8] "~$signment .doc"   
## [9] "¶ºÓÎÓÎÏ·ºÐ.lnk"   
## [10] "2020-11-10-Take-Home-Exam-Part-3.B.DiD.html"   
## [11] "360Ãâ·ÑWiFi.lnk"   
## [12] "ceosal2.RData"   
## [13] "Chinese manufacturing on the move Factor supplyor market access.pdf"   
## [14] "CHIP2013 Questionaire.rar"   
## [15] "CV Chinese Version.docx"   
## [16] "data\_r.csv"   
## [17] "desktop.ini"   
## [18] "Discord.lnk"   
## [19] "Does hukou still matter The household registration system and its impact on social stratification and mobility in China.pdf"  
## [20] "Dota 2.url"   
## [21] "Economics-Lab07.docx"   
## [22] "Final exam"   
## [23] "Final exam partA.docx"   
## [24] "Final exam.zip"   
## [25] "First meeting"   
## [26] "HUMAN-CAPITAL EXTERNALITIES IN CHINA.pdf"   
## [27] "Lab-01.docx"   
## [28] "Lab-03.Rmd"   
## [29] "Lab 01.Rmd"   
## [30] "Lab 03.Rmd"   
## [31] "Lab 10 for ECO R002.docx"   
## [32] "Microsoft Teams.lnk"   
## [33] "Ñ§ÀúÈÏÖ¤"   
## [34] "ÕË»§¼°ÃÜÂë.txt"   
## [35] "PhD Econometrics"   
## [36] "Population aged 6 and over by sex, education attainment and region.xls"   
## [37] "RStudio.lnk"   
## [38] "Ruby"   
## [39] "sales-data (1).csv"   
## [40] "sales-data.csv"   
## [41] "Software-R.zip"   
## [42] "Third Meeting Note.docx"   
## [43] "UEA PhD School File"   
## [44] "UK"   
## [45] "VISA"   
## [46] "Wallpaper Engine.url"   
## [47] "WhatsApp.lnk"

## This is a “.csv” file. In order to load a csv file into R we need to use the function “read.csv()”. When using this function we must name the data while loading it.

sales <- read.csv("sales-data.csv")

## convert the data into the data.table format

dt.sales <- data.table(sales)  
rm(sales)

## Explore the data

## Review some Lab01 steps

ncol(dt.sales)

## [1] 2

nrow(dt.sales)

## [1] 22

colnames(dt.sales)

## [1] "sales" "advertising"

head(dt.sales)

## sales advertising  
## 1: 999 48  
## 2: 1169 50  
## 3: 1036 68  
## 4: 643 52  
## 5: 988 76  
## 6: 1076 74

stargazer(dt.sales, type = "text")

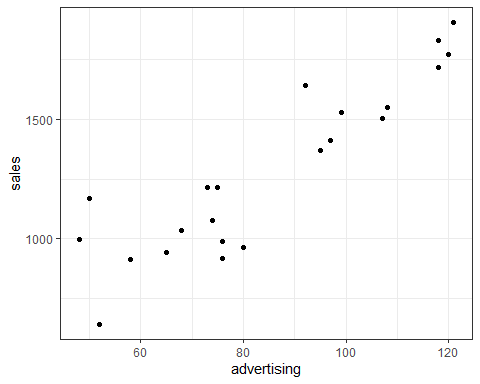
##   
## =============================================================  
## Statistic N Mean St. Dev. Min Pctl(25) Pctl(75) Max   
## -------------------------------------------------------------  
## sales 22 1,286.636 353.621 643 990.8 1,543.8 1,905  
## advertising 22 85.000 23.759 48 69.2 105 121   
## -------------------------------------------------------------

summary(dt.sales)

## sales advertising   
## Min. : 643.0 Min. : 48.00   
## 1st Qu.: 990.8 1st Qu.: 69.25   
## Median :1215.0 Median : 78.00   
## Mean :1286.6 Mean : 85.00   
## 3rd Qu.:1543.8 3rd Qu.:105.00   
## Max. :1905.0 Max. :121.00

## And we can use plots to explore the data. In this case, we are dealing with two continuous variables so it makes sense to use a scatter plot.

qplot( data = dt.sales  
, x = advertising  
, y = sales  
, geom = "point") +  
theme\_bw()

 ## What relationship do we observe? In this case, we can see that the two variables are positively correlated.We can use the function “cor” to get the exact correlation coefficient.

dt.sales[, cor(sales, advertising)]

## [1] 0.9003409

## In order to know whether the correlation coefficient is statistically significant we can use:

dt.sales[, rcorr(sales, advertising)]

## x y  
## x 1.0 0.9  
## y 0.9 1.0  
##   
## n= 22   
##   
##   
## P  
## x y   
## x 0  
## y 0

## Simple Regression Analysis

## In order to know what is the change in sales that we can expect from increasing our advertising investment by one dollar we can create a simple regression model.

lm.sales <- lm(sales ~ advertising, data=dt.sales)

## In order to get the coefficient estimates, significance levels, and the measures for the quality of our model, we use the “summary” function.

summary(lm.sales)

##   
## Call:  
## lm(formula = sales ~ advertising, data = dt.sales)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -254.63 -71.78 -17.34 82.97 351.38   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 147.590 127.618 1.157 0.261   
## advertising 13.401 1.448 9.252 1.15e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 157.7 on 20 degrees of freedom  
## Multiple R-squared: 0.8106, Adjusted R-squared: 0.8011   
## F-statistic: 85.6 on 1 and 20 DF, p-value: 1.15e-08

stargazer(lm.sales, type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## sales   
## -----------------------------------------------  
## advertising 13.401\*\*\*   
## (1.448)   
##   
## Constant 147.590   
## (127.618)   
##   
## -----------------------------------------------  
## Observations 22   
## R2 0.811   
## Adjusted R2 0.801   
## Residual Std. Error 157.691 (df = 20)   
## F Statistic 85.604\*\*\* (df = 1; 20)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## Extract the parameters of the estimated regression equation using the coefficients function

coeffs = coefficients(lm.sales)  
coeffs

## (Intercept) advertising   
## 147.59047 13.40054

## Interpretation

##β0 = 147.6 gives us the average sales level when the advertising investment is zero. ##β1 = 13.4 gives us the increase in sales that results from a 1 unit (dollar) increase in advertising investment. ##R2 gives us the percentage of the variation in sales that is explained by the variation in the advertising investment

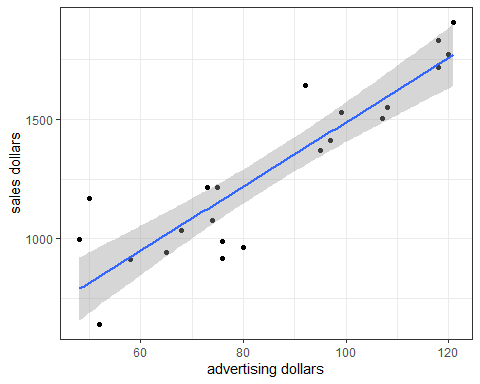
## Plot

## Plot the relationship between advertising and sales now also plotting the regression line.

qplot( data = dt.sales  
, x = advertising  
, y = sales  
, geom = c("point", "smooth")  
, method = lm) +  
theme\_bw() +  
labs( x = "advertising dollars", y = "sales dollars")

## Warning: Ignoring unknown parameters: method

## `geom\_smooth()` using formula 'y ~ x'

 ## Predicted values ## Obtain the predicted sales for an advertising investment of 100

advertising = 100  
sales = coeffs[1] + coeffs[2]\*advertising  
sales

## (Intercept)   
## 1487.644

## Alternatively, you can use the “predict” function to do this automatically. We first wrap the parameters inside a new data table variable called newdata.

my.budget = data.table(advertising=100)

## We then apply the predict function and set the predictor variable in the newdata argument. We also set then interval type as “predict”, and use the default 0.95 confidence level.

predict(lm.sales, my.budget)

## 1   
## 1487.644

predict(lm.sales, my.budget, interval="predict")

## fit lwr upr  
## 1 1487.644 1148.274 1827.014

## Hypothesis Testing

## Steps

# 1. State the hypotheses

# 2. Identify the appropriate test statistic and its distribution

# 3. Specify the significance level

# 4. State the decision rule - critical value(s)

# 5. Collect the data and compute the test statistic

# 6. Make the statistical decision

# 7. Make the practical decision

## Example1

## Hypothesis Testing for a Population Mean with known variance

alpha = .05 # significance level  
z.half.alpha = qnorm(1-alpha/2)  
c(-z.half.alpha, z.half.alpha) # critical values

## [1] -1.959964 1.959964

Thus, the decision rule is that the null hypothesis should be rejected if z <= −1.96 or z >= 1.96. Compute the test statistic:

xbar = 1000/25 # sample mean  
mu0 = 45 # hypothesized value  
sigma = sqrt(500) # population standard deviation  
n = 25 # sample size  
z = (xbar-mu0)/(sigma/sqrt(n)) # test statistic  
z

## [1] -1.118034

## we can apply the pnorm function to compute the two-tailed p-value of the test statistic.

pnorm(z, lower.tail=FALSE) # upper tail

## [1] 0.8682238

pnorm(z, lower.tail=TRUE) # lower tail

## [1] 0.1317762

## For the two-tailed p-value we choose the minimum of these values, which in this case is the lower tail, which happens to be the default option in R.

pval = 2 \* pnorm(z) # lower tail  
pval # two-tailed p-value

## [1] 0.2635525

## Hypothesis Testing for a Population Mean with unknown variance

t.alpha = .05 # significance level  
t.half.alpha = qt(1-alpha/2, 25-1)  
c(-t.half.alpha, t.half.alpha) # critical values

## [1] -2.063899 2.063899

## Thus, the decision rule is that the null hypothesis should be rejected if t <= −2.064 or t >= 2.064.

## Compute the test statistic:

xbar = 1000/25 # sample mean  
mu0 = 45 # hypothesized value  
s = sqrt(400) # sample standard deviation  
n = 25 # sample size  
t = (xbar-mu0)/(s/sqrt(n)) # test statistic  
t

## [1] -1.25

## We are not able to reject the null.

# Instead of using the critical value, we can apply the pt function to compute the two-tailed p-value of the test statistic.

pt(t, df=25-1, lower.tail=FALSE) # upper tail

## [1] 0.8883243

pt(t, df=25-1, lower.tail=TRUE) # lower tail

## [1] 0.1116757

## For the two-tailed p-value we choose the minimum of these values, which in this case is the lower tail, which happens to be the default option in R.

pval = 2 \* pt(t, df=25-1) # lower tail  
pval # two-tailed p-value

## [1] 0.2233515

## Since it turns out to be greater than the .05 significance level, we do not reject the null hypothesis.

## Hypothesis Testing for a Population Variance

## Are there reasons to doubt the value of the variance?

## The critical values at 0.05 significance level are:

q.alpha = .05 # significance level  
q.half.alpha.up = qchisq(1-alpha/2, 25-1) # critical values  
q.half.alpha.up

## [1] 39.36408

q.half.alpha.low = qchisq(alpha/2, 25-1) # critical values  
q.half.alpha.low

## [1] 12.40115

## Thus, the decision rule is that the null hypothesis should be rejected if Q <= 12.4 or Q >= 39.4.

## Compute the test statistic

sigma\_sqr\_0 = 500 # hypothesized value  
s\_sqr = 400 # sample standard deviation  
n = 25 # sample size  
Q = ((n-1)\*s\_sqr)/sigma\_sqr\_0 # test statistic  
Q

## [1] 19.2

## Example 2:

## Hypothesis Testing for a Proportion

## A particular type of cancer therapy has a 60% success rate. A group of researchers developed a new type of treatment and its effectiveness is to be tested. In 61 cases, 47 were successfully treated.

## Is there enough empirical evidence that allows us to conclude that the new treatment is better than the old

one? Use α = 0.05.

alpha = .05 # significance level  
z.alpha = qnorm(1-alpha)  
z.alpha # critical value

## [1] 1.644854

## Thus, the decision rule is that the null hypothesis should be rejected if z >= 1.65.

## Compute the test statistic

p0 = 0.60  
fn = 47/61 # sample proportion  
n = 61 # sample size  
z = (fn-p0)/sqrt((0.6\*(1-0.6))/n) # test statistic  
z

## [1] 2.718084

## We reject the null.

## Example 3

##Hypothesis Testing for a Difference in Population Means - Independent Samples and Variance Unknown

## Is there a difference between average dividends of the stocks in Dow Jones and the ones in Eurostoxx, knowing they have equal variances and normally distributed? The data of two independent samples is the following, use α = 0.05.

nX = 21  
nY = 25  
sX = 1.30  
sY = 1.16  
Sp\_sqr = ((nX - 1)\*(sX^2) + (nY-1)\*(sY^2))/(nX + nY - 2)  
Sp\_sqr

## [1] 1.502145

df = nX + nY - 2  
df

## [1] 44

##The critical values at 0.05 significance level are:

alpha = .05 # significance level  
t.half.alpha = qt(1-alpha/2, 44)  
c(-t.half.alpha, t.half.alpha) # critical values

## [1] -2.015368 2.015368

## Thus, the decision rule is that the null hypothesis should be rejected if Q <= −2.02 or z >= 2.02.

## Compute the test statistic

xbar = 3.27  
ybar = 2.53  
t = ((xbar-ybar)-0)/sqrt((Sp\_sqr/nX)+(Sp\_sqr/nY))  
t

## [1] 2.039748

## We reject the null.

## Instead of using the critical value, we can apply the pt function to compute the two-tailed p-value of the test statistic:

pt(t, df=44, lower.tail=FALSE) # upper tail

## [1] 0.02370372

pt(t, df=44, lower.tail=TRUE) # lower tail

## [1] 0.9762963

## For the two-tailed p-value we choose the minimum of these values, which in this case is the upper tail.

pval = 2 \* pt(t, df=44, lower.tail=FALSE)  
pval # two-tailed p-value

## [1] 0.04740744

## Since it turns out to be smaller than the .05 significance level, we reject the null hypothesis.

## Example 4

# change the file’s path to your own  
dt.stocks <- data.table(read.csv("data\_r.csv"))  
dt.stocks <- setnames(dt.stocks, tolower(names(dt.stocks)))  
head(dt.stocks)

## serial year month djcomp djind djutil djtran nasdaq sp500 sp100  
## 1: 1 1990 Jan 959.54 2590.54 223.65 1045.87 415.8 329.08 307.88  
## 2: 2 1990 Feb 986.07 2627.25 220.38 1129.09 425.8 331.89 312.48  
## 3: 3 1990 Mar 1012.10 2707.21 214.66 1183.14 435.5 339.94 320.03  
## 4: 4 1990 Apr 979.70 2656.76 203.09 1129.98 420.1 330.80 314.23  
## 5: 5 1990 May 1040.16 2876.66 211.39 1171.53 459.0 361.23 342.66  
## 6: 6 1990 Jun 1031.07 2880.69 210.01 1142.70 462.3 358.02 339.80  
## treas3m idjcomp idjind idjutil idjtran inasdaq isp500 isp100 itreas3m  
## 1: 7.90 NA NA NA NA NA NA NA NA  
## 2: 8.00 2.76 1.42 -1.46 7.96 2.41 0.85 1.49 0.64  
## 3: 8.17 2.64 3.04 -2.60 4.79 2.28 2.43 2.42 0.66  
## 4: 8.04 -3.20 -1.86 -5.39 -4.49 -3.54 -2.69 -1.81 0.65  
## 5: 8.01 6.17 8.28 4.09 3.68 9.26 9.20 9.05 0.64  
## 6: 7.99 -0.87 0.14 -0.65 -2.46 0.72 -0.89 -0.83 0.64

## Confidence Intervals

## Calculate the 95% confidence interval for the stocks’ means. Example:

xbar <- dt.stocks[, mean(idjcomp, na.rm=TRUE)]  
s <- dt.stocks[, sd(idjcomp, na.rm=TRUE)]  
n <- dt.stocks[, length(which(!is.na(idjcomp)))]  
error <- qnorm(0.975)\*s/sqrt(n)  
left <- xbar-error  
right <- xbar+error  
left

## [1] 0.2391836

right

## [1] 1.156545

## Inference for the population mean.

## Lets look at some examples using our stock data. For instance, if we wanted to check whether the mean of S&P500 is equal to zero we would write:

dt.stocks[, t.test(isp500)]

##   
## One Sample t-test  
##   
## data: isp500  
## t = 2.8631, df = 294, p-value = 0.004497  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 0.220257 1.188896  
## sample estimates:  
## mean of x   
## 0.7045763

## If we wanted to test whether it is greater than some specific value (say 0.5) and use a 99% confidence interval, we would write:

dt.stocks[, t.test(isp500, alternative = c("greater"), mu=0.5, conf.level = 0.99)]

##   
## One Sample t-test  
##   
## data: isp500  
## t = 0.83131, df = 294, p-value = 0.2032  
## alternative hypothesis: true mean is greater than 0.5  
## 99 percent confidence interval:  
## 0.1289499 Inf  
## sample estimates:  
## mean of x   
## 0.7045763

## Inference for difference of population means - paired samples

## Paired t-test: t.test(y1,y2,paired=TRUE) where y1 & y2 are numeric.

## Say we want to compare between IDJCOMP and INASDAQ.

dt.stocks[, t.test(idjcomp, inasdaq, paired=TRUE)]

##   
## Paired t-test  
##   
## data: idjcomp and inasdaq  
## t = -1.178, df = 294, p-value = 0.2397  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.9038613 0.2269799  
## sample estimates:  
## mean of the differences   
## -0.3384407

## Inference for difference of population means - independent samples

## Impact of crisis on stock indices

## Create an indicator variable that takes the value of 1 if it is post-2008 (year of the economic crisis)

dt.stocks[, postcrisis:=ifelse(year>2008,1,0)]

## You can use the indicator variable to look at the mean value of the stock variation before and after the crisis

dt.stocks[postcrisis==0, mean(idjcomp, na.rm=TRUE)]

## [1] 0.5946696

dt.stocks[postcrisis==1, mean(idjcomp, na.rm=TRUE)]

## [1] 1.042353

## Then you can use the t.test to check whether the difference in means before and after the crisis is statistically significant

dt.stocks[, t.test(idjcomp ~ postcrisis)]

##   
## Welch Two Sample t-test  
##   
## data: idjcomp by postcrisis  
## t = -0.77236, df = 103.8, p-value = 0.4417  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.5971454 0.7017787  
## sample estimates:  
## mean in group 0 mean in group 1   
## 0.5946696 1.0423529

## You can also use the t-test to compare between the means of two different variables.

## Independent 2-group t-test: t.test(y1,y2) where y1 and y2 are numeric.

dt.stocks[, t.test(idjcomp, inasdaq, var.equal=TRUE)]

##   
## Two Sample t-test  
##   
## data: idjcomp and inasdaq  
## t = -0.75383, df = 588, p-value = 0.4513  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.2202060 0.5433246  
## sample estimates:  
## mean of x mean of y   
## 0.6978644 1.0363051

dt.stocks[, t.test(idjcomp, inasdaq, var.equal=FALSE)]

##   
## Welch Two Sample t-test  
##   
## data: idjcomp and inasdaq  
## t = -0.75383, df = 486.57, p-value = 0.4513  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.2205853 0.5437039  
## sample estimates:  
## mean of x mean of y   
## 0.6978644 1.0363051