Homework 2 - Statistical Modeling (Master Degree in Data Science)

Matteo Suardi

2025-04-26

Practical part

The file `energy_efficiency.Rdata` contains simulated data related to home energy efficiency improvements. The dataset records the amount of money invested in energy-saving upgrades (variable `investment`, measured in thousands of Euros) and the corresponding average monthly electricity consumption of the homes after the upgrades (variable `consumption`, measured in kilowatt-hours, kWh).

1. Describe the observed values. Which is the correlation?

Let's load the dataset first:

```
load("energy_efficiency.Rdata")
View(energy)
```

Dimensions are:

```
dim(energy)
```

```
## [1] 187   2
```

```
head(energy)
```

```
##   investment consumption
## 1  9.4186451    32.75020
## 2  2.2447415    33.16612
## 3  0.6566587    35.06405
## 4  1.1829064    34.59358
## 5  9.6640546    32.43188
## 6  6.2462955    33.26537
```

The dataset consists of 187 observations and 2 variables: `investment` and `consumption`.

```
require(skimr)
```

```
## Loading required package: skimr
```

```r
skim_without_charts(energy)
```

Table 1: Data summary

Name	energy
Number of rows	187
Number of columns	2
Column type frequency:	
numeric	2
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
investment	0	1	5.21	3.04	0.03	2.64	5.42	8.04	9.94
consumption	0	1	33.32	1.73	28.22	32.18	33.43	34.69	37.98

There are no missing values.

The distribution of `investment` appears approximately symmetric, centered around 5.4, with no evidence of strong skewness. The mean (5.21) and median (5.42) are close, and the interquartile range is balanced.

The distribution of `consumption` is tightly concentrated around the mean and approximately symmetric, with a very low standard deviation (1.73). The mean (33.32) and median (33.43) are nearly identical, and only a mild left tail is observed.

```r
round(cor(energy$investment, energy$consumption), 2)
```

```
## [1] -0.54
```

The Pearson correlation coefficient between the two variables is -0.54.

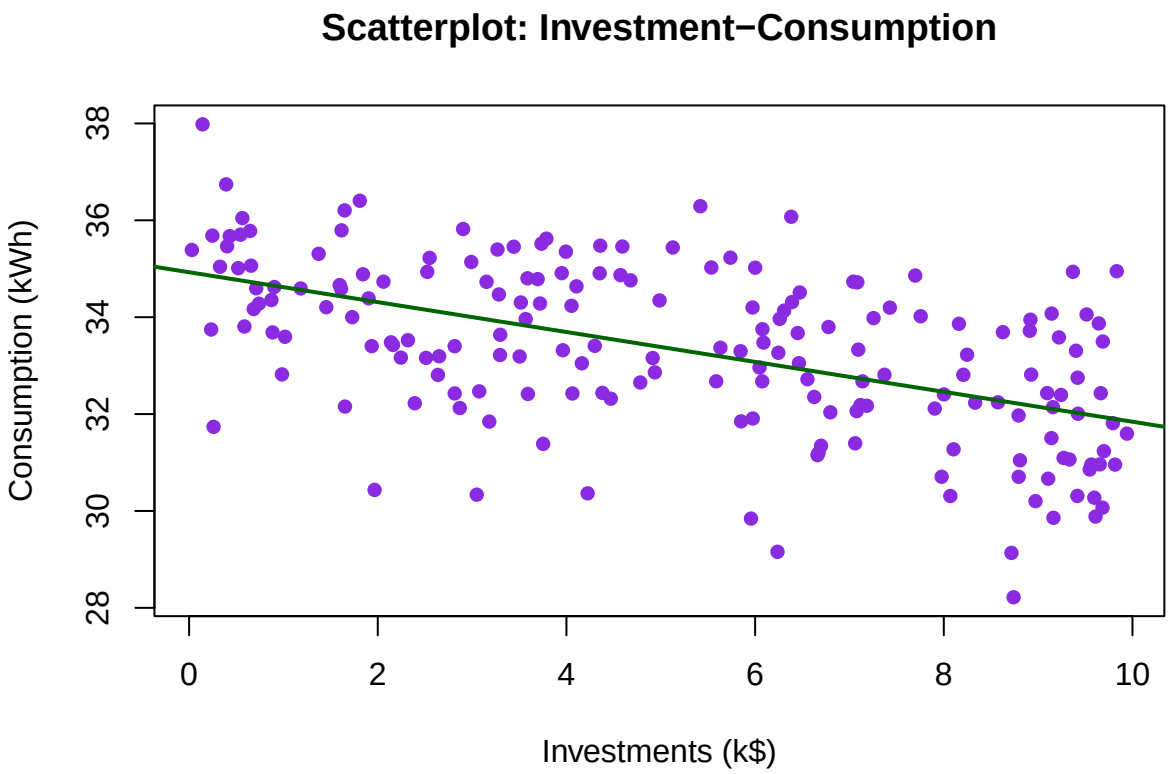
This value indicates a **moderate negative linear association** between investments and consumption: homes that invested more in energy-saving upgrades tend to have lower electricity consumption.

However, the correlation is not extremely strong, implying that while investment plays a role in reducing consumption, other unobserved factors might also influence energy efficiency.

2. Present some descriptive plots to show the data and comment on them.

Scatter plot

```r
plot(energy$investment, energy$consumption,
     pch = 16, col = "blueviolet",
     xlab = "Investments (k$)",
     ylab = "Consumption (kWh)",
     main = "Scatterplot: Investment-Consumption")
abline(lm(consumption ~ investment, data = energy), col = "darkgreen", lwd = 2)
```



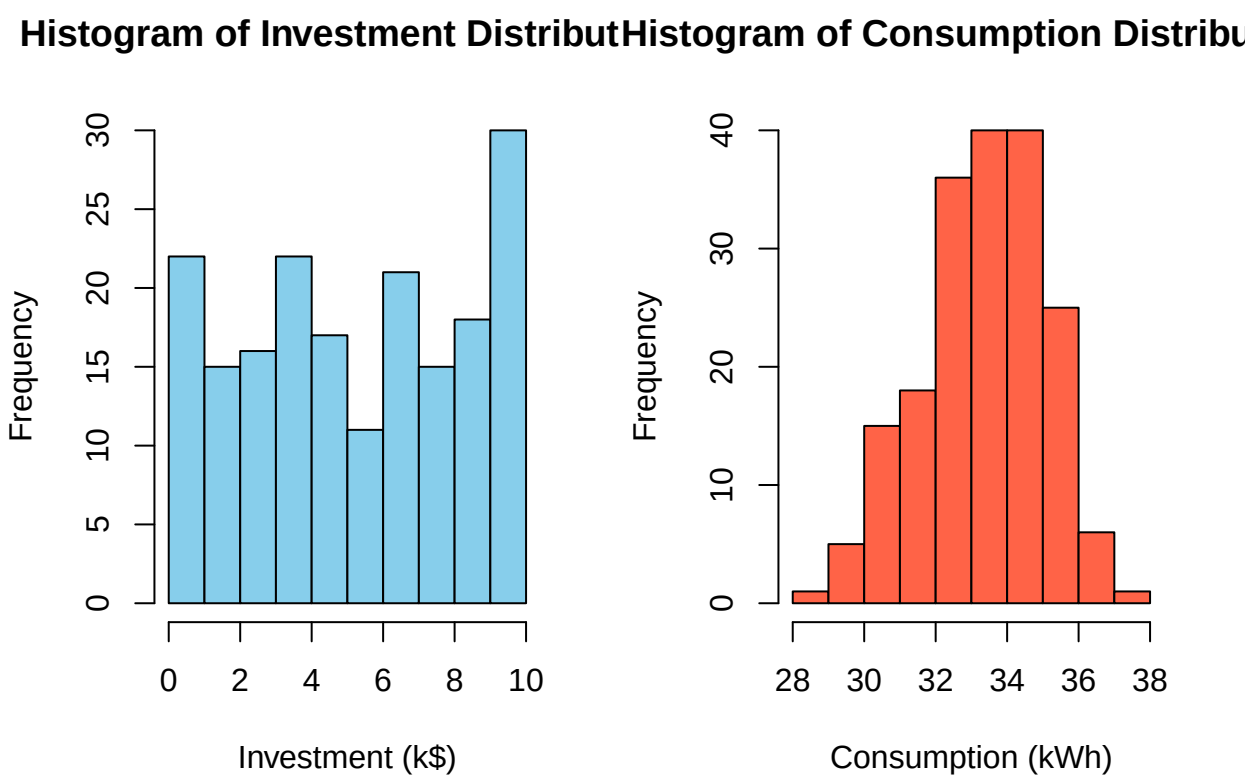
The scatterplot clearly illustrates the inverse relationship between investment and consumption. This indicates (shown by the superimposed dark green regression line) **negative trend**, which confirms the previously computed correlation of -0.54.

Each point represents a household, and we observe that homes which invested more in energy-saving upgrades generally display lower post-upgrade electricity consumption. While the trend is evident, the spread of the points (particularly in the mid-range investments levels) suggests that other unobserved factors may also influence consumption, and the relationship is not perfectly deterministic.

Additionally, the presence of few high-consumption households with low investments may represent non-responders or homes where upgrades were insufficiently effective.

Histograms

```
par(mfrow = c(1,2))
hist(energy$investment,
     col = "skyblue",
     main = "Histogram of Investment Distribution",
     xlab = "Investment (k$)")
hist(energy$consumption,
     col = "tomato",
     main = "Histogram of Consumption Distribution",
     xlab = "Consumption (kWh)")
```



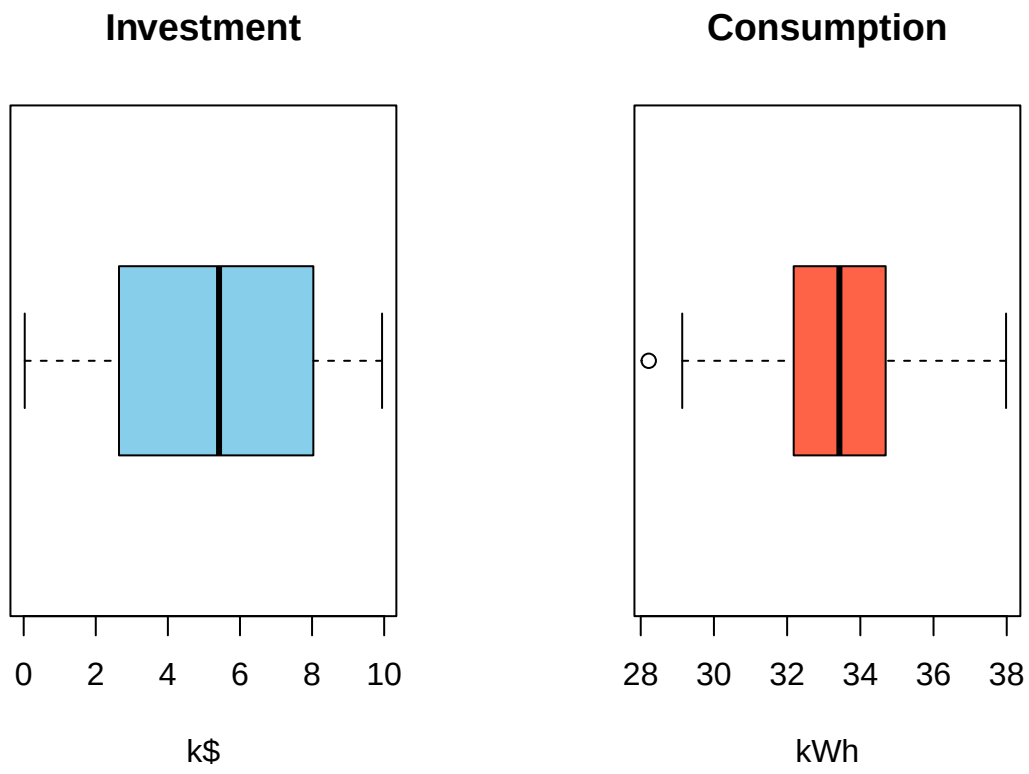
The `investment` variable appears fairly uniform, with values distributed across the full range from 0 to 10 k$. There's no clear mode, and the distribution does not show strong skewness, although a slightly higher frequency is observed at the upper end for higher amounts of investments.

The `consumption` variable shows a unimodal distribution centered around 33-34 kWh. However, it is not perfectly symmetric: there's a slight left skewness, as the lower tail (28-31 kWh) extends further than the upper tail.

This suggests the presence of a small number of homes with particularly low consumption even after energy-saving investments.

Box plots

```
par(mfrow = c(1,2))
boxplot(energy$investment,
        horizontal = TRUE,
        col = "skyblue",
        xlab = "k$",
        main = "Investment")
boxplot(energy$consumption,
        horizontal = TRUE,
        col = "tomato",
        xlab = "kWh",
        main = "Consumption")
```



The box plot of `investment` shows a symmetric and well-spread distribution, without apparent outliers. The central 50% of values (interquartile range) covers a wide portion of the investment scale, from approximately 2.6 to 8 k$.

The `consumption` box plot reveals a more compact distribution, with most values concentrated between 32 and 35 kWh. However, a single low outlier appears below the whiskers, indicating the presence of a home with exceptionally low electricity usage, possibly due to particularly effective upgrades or additional unobserved factors.

The presence of outliers also helps explain the left skewness observed in the histogram, and should be taken into account in further analysis.

3. We are interested in providing a measure of uncertainty for the estimated correlation. Apply the bootstrap procedure and describe the results. We aim to assess the uncertainty of the sample correlation coefficient $\hat{\rho}$ between investment and consumption. Since the theoretical distribution of the statistic is unknown, we apply the non-parametric bootstrap method.

First, let's generate bootstrap samples.

```
set.seed(16253)
B <- 1000 # academic standard
n <- nrow(energy)
Tboot <- rep(0, B)

for (i in 1:B) {
  idx <- sample(1:n, n, replace = TRUE)
  Xstar <- energy[idx, ]
  Tboot[i] <- cor(Xstar$investment, Xstar$consumption)
```

```
}

round(sd(Tboot), 2)
```

```
## [1] 0.05
```

We followed the standard non-parametric bootstrap procedure: random resampling with replacement from the original data and computation of the statistic of interest (`correlation`) on each resampled dataset.

Let's compare this manual implementation with `bootstrap` library.

```
require(bootstrap)
```

```
## Loading required package: bootstrap
```

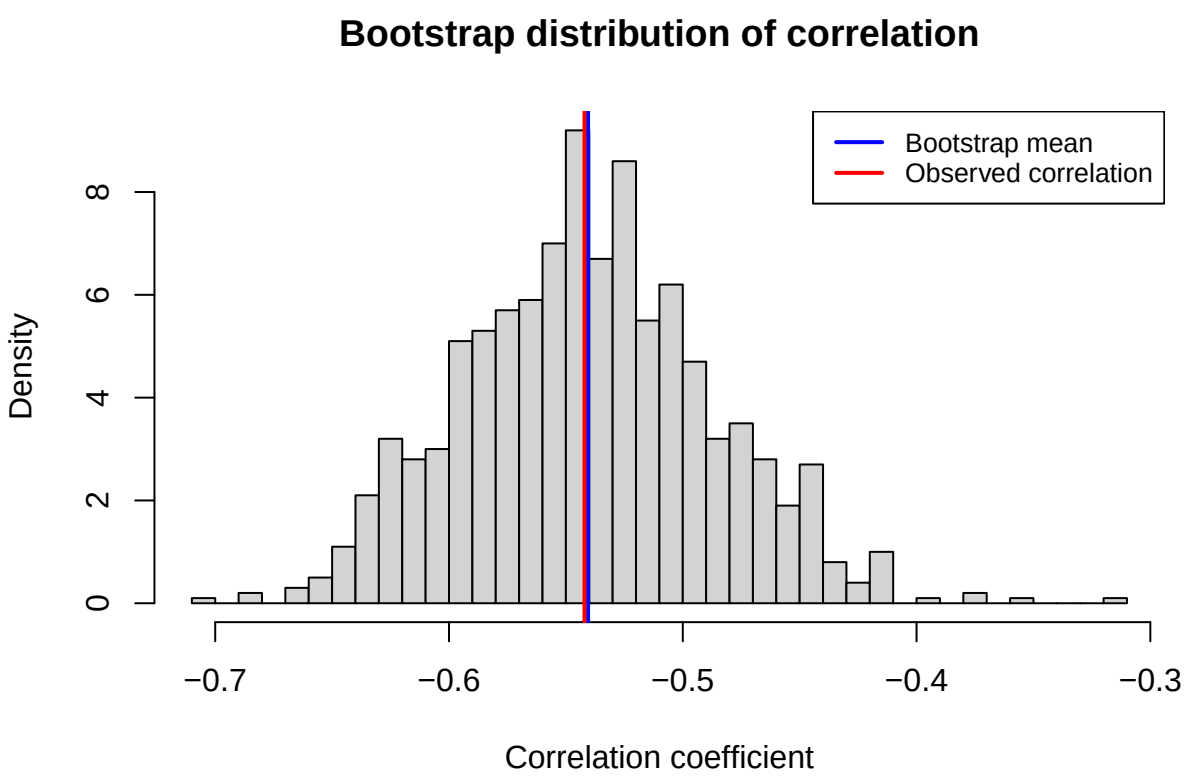
```
corr_fun <- function(z) {
  d <- matrix(z, ncol = 2, byrow = FALSE)
  cor(d[,1], d[,2])
}

bootM <- bootstrap(cbind(energy$investment, energy$consumption),
                   nboot = 1000,
                   corr_fun)

round(sd(bootM$thetastar), 2)
```

```
## [1] 0.07
```

When comparing the manual bootstrap procedure with the `bootstrap` library, a small difference is observed in the standard errors (0.05 vs 0.07).

This is because the manual bootstrap correctly resamples entire observations (preserving the natural association between investment and consumption), while the `bootstrap()` function resamples data without maintaining the original pairing structure, thus introducing additional variability.

```
hist(Tboot, breaks = 50,
     col = "lightgray", freq = FALSE,
     main = "Bootstrap distribution of correlation",
     xlab = "Correlation coefficient")
abline(v = mean(Tboot), col = "blue", lwd = 2)
abline(v = cor(energy$investment, energy$consumption),
       col = "red",
       lwd = 2)
legend("topright",
       legend = c("Bootstrap mean", "Observed correlation"),
       col = c("blue", "red"),
       lty = 1,
       lwd = 2,
       cex = 0.8)
```



The histogram shows the bootstrap distribution of the correlation coefficient based on 1000 bootstrap samples. The distribution is approximately centered around the observed correlation, with the bootstrap mean (blue line) closely matching the observed sample correlation (red line).

The shape is moderately symmetric, but with a slightly heavier right tail: a few bootstrap samples produced correlation estimates closer to -0.3, indicating the presence of mild extreme values.

This suggests that although the bootstrap correlation is generally stable, occasional large deviations from the mean can occur.

The presence of these mild outliers justifies the careful interpretation of the bootstrap confidence intervals and reinforces the importance of using robust resampling techniques to quantify uncertainty.

4. Report also the confidence interval at 99% confidence level. Comment on the result.

We compute the 99% bootstrap percentile confidence interval for the correlation coefficient. The use of the percentile method is particularly appropriate in this context since the bootstrap distribution of the correlation appears approximately symmetric and unimodal, with only mild right skewness.

```
round(quantile(Tboot, c(0.005, 0.995)), 2)
```

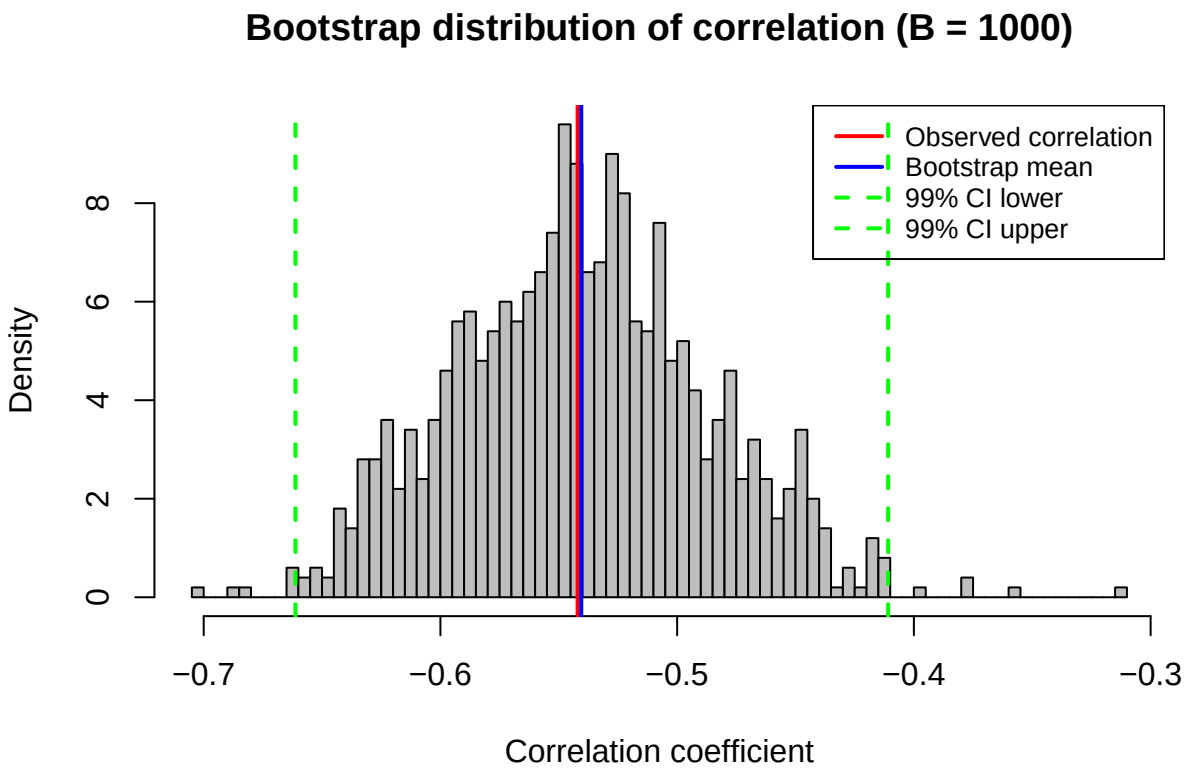
```
##  0.5% 99.5%
## -0.66 -0.41
```

The obtained result is:

$$[-0.66, -0.421]$$

Based on the bootstrap procedure, we are 99% confident that the true correlation coefficient between investment and consumption falls within this range. Recall, however, that in frequentist terms, the confidence level refers to the long-run proportion of intervals containing the true parameter, not to the probability that the specific observed interval contains the parameter. The interval is entirely negative, confirming the existence of a significant negative association between the two variables.

```r
sB <- mean(Tboot)
Q <- quantile(Tboot, c(0.005, 0.995))

hist(Tboot,
     breaks = 60,
     freq = FALSE,
     main = "Bootstrap distribution of correlation (B = 1000)",
     xlab = "Correlation coefficient",
     col = "gray",
     ylab = "Density")

abline(v = cor(energy$investment, energy$consumption), col = "red", lwd = 2)
abline(v = sB, col = "blue", lwd = 2)
abline(v = Q[1], col = "green", lwd = 2, lty = 2)
abline(v = Q[2], col = "green", lwd = 2, lty = 2)

legend("topright",
       legend = c("Observed correlation", "Bootstrap mean", "99% CI lower", "99% CI upper"),
       col = c("red", "blue", "green", "green"),
       lty = c(1, 1, 2, 2),
       lwd = 2,
       cex = 0.8)
```



The green dashed lines represent the 99% percentile-based confidence interval $[-0.66, -0.41]$.

As previously said, the interval is entirely negative, reinforcing the evidence of a significant negative linear relationship between investment and electricity consumption after energy-efficiency upgrades.

The concentration of bootstrap replications around the observed correlation, combined with a moderate spread and the presence of some mild extreme values, indicates that while there is some natural sampling variability, the negative association is a robust and reliable result.

Overall, the bootstrap distribution provides an effective and intuitive way to assess the uncertainty associated with the correlation estimate.

5. Describe the bootstrap distribution and the estimated values. The bootstrap distribution of the correlation coefficient appears approximately symmetric, centered around the observed correlation value (-0.54).

However, it shows a slight right skewness, with a few replications producing weaker (less negative) correlation estimates (up to about -0.3).

The bulk of the distribution is concentrated between approximately -0.66 and -0.42.

The shape suggests that the bootstrap procedure produces stable and consistent estimates of the true correlation.

The estimated values are:

- **Observed sample correlation**: $\hat{\rho}_{\text{obs}} = -0.54$
- **Bootstrap mean** (average of bootstrap replications):

$$\text{mean}(T_{\text{boot}}) \approx -0.541$$

- **Bootstrap standard error** (variability of the bootstrap replications):

$$\mathrm{SE}_{\mathrm{boot}} = \mathrm{sd}(\mathrm{T}_{\mathrm{boot}}) \approx 0.05$$

- **99% bootstrap percentile confidence interval**: $[-0.66, -0.42]$

The bootstrap mean is extremely close to the observed sample correlation, indicating a negligible bias introduced by the resampling process.

The bootstrap standard error is small relative to the range of possible correlation values $[-1, 1]$ and to the observed variability in the data, suggesting that the correlation estimate is relatively precise.

The concentration of bootstrap estimates, with only mild deviations, reinforces the robustness and reliability of the negative association identified between investment and electricity consumption.

In conclusion, the bootstrap procedure provides strong support for the stability of the estimated correlation, with minimal bias and moderate uncertainty.

Theoretical part

Consider the following statements concerning bootstrap, and indicate if they are true or false. If false, correct them.

- *Bootstrap is a resampling method that involves creating many random samples without replacement from an observed dataset, in order to estimate the distribution of a statistic.*

False. Bootstrap is a resampling method that involves creating many random samples **with replacement** from an observed dataset, in order to estimate the distribution of a statistic.

- *The theoretical distribution of the statistic is complex or unknown.*

True.

- *The sample is small and strong parametric assumptions are not desirable.*

True.

- *One wants to estimate confidence intervals or the variability of a parameter (e.g., mean, median, regression coefficients, etc.).*

True.

- *It does not require strong assumptions about the underlying data distribution.*

True.

- *It is flexible and applicable to many different statistics.*

True.

- *It can be computationally intensive.*

True.

- *If the original sample is strongly biased, the bootstrap will not reflect that bias.*

False. If the original sample is strongly biased, the bootstrap will reflect and reproduce that bias.