University of Amsterdam
Economics & Business

Advanced Econometrics II: Assignment on Bootstrap

Alessandro Bonetti	Bowen Ma	Olga Milovanova
15855902	12960780	12879436

January 13, 2025

Declaration of Originality

1. These solutions are solely our own work.

2. We have not made (part of) these solutions available to any other student.

3. We shall not engage in any other activities that will dishonestly improve our results or dishonestly improve or hurt the results of others.

Question 1

The observations of Figure 2 for different θ are summarized below.

- **For $\boldsymbol{\theta = 0}$:** At 0 there is a substantial mass in the histogram, this is because, when the true θ equals 0, the sample mean can also become negative (approximately half of the time), which will automatically become 0. As a result, the distribution is skewed and has a point mass at 0.
- **For $\boldsymbol{\theta = \{0.1, 0.2, 0, 3\}}$:** For θ slightly above 0 the probability of the sample mean being below 0 decreases. Consequently, fewer estimates equal 0, although there might still be a noticeable fraction of estimates at 0 when θ is small, especially for $\theta = 0.1$.
- **For $\boldsymbol{\theta = \{0.4, 0.5, 0, 6\}}$:** As θ grows, the sample mean is more likely to be positive, and hence even fewer estimates equal 0. For sufficiently large θ, the distribution of the estimate begins to look similar to a normal distribution (centered around θ) with minimal mass at 0.

Question 2

For this question, we have considered several tests for normality. The two tests that performed the best were the Anderson-Darling and the Shapiro-Wilk tests. This is expected due to the following reasons:

- **Shapiro-Wilk** is frequently cited as having good overall power for a wide range of alternative distributions, particularly with smaller sample size.
- **Anderson-Darling** tends to be powerful for detecting tail issues, which is a good approach here due to the plots in question 1, where for smaller θ the distribution tends to have heavier tails.

From Figure 3 we see that both tests reject normality for smaller θ the strongest for $n = 100$. For a larger sample size ($n = 400$), the Shapiro-Wilk test is still the most powerful, although Doornik-Hansen performs very similar to Anderson-Darling. For higher θ the tests show similar results, with Anderson-Darling test slightly under-performing for $n = 400$. Concluding, the test with the overall best performance in rejecting the normality for smaller θ is the Shapiro-Wilk test. We will use this test for the remainder of this assignment.

Question 3

In order to investigate the non-normality of the sample mean $\hat{\theta}_n$ Q-Q plots were used. In Figure 4 the plots are depicted for sample sizes $n \in \{100, 200, 300\}$ and a t-distribution with $df \in \{1, 2, 5\}$ was used to generate the data. The 45 degree line, which is indicated by the red color, suggests that the data follows a normal distribution. It can be observed that:

- **For $\boldsymbol{df = 1}$:** the distribution deviates from the red line (the normal distribution) the most. The part of the distribution that shows the largest deviation from normality are the heavy-tails of $\hat{\theta}_n$. These are the points far left and far right of the Q-Q plot that deviate the most from the red line.
- **For $\boldsymbol{df = 2}$:** the distribution still deviates from the normal distribution, although visually the distribution is less non-normal. In this case large outliers are still present.
- **For $\boldsymbol{df = 5}$:** the data shows the least deviation from normality.

The first two cases deviate from the normal distribution the most due to infinite variance. In the last case the variance is finite, showing the most normal results. There is no significant difference in results between the chosen sample sizes.

Question 4

For this question, the Shapiro-Wilk test has been applied in order to test for normality. The simulations were conducted for several degrees of freedom $df \in \{1, 2, 5\}$ of the t-distribution, sample sizes $n \in \{100, 200, 300\}$ and variations in bootstrap sample sizes $m \in \{n^{1/2}, n^{4/5}, \frac{n}{2}, n\}$.

In Figure 5 the results are presented. It can be observed that for larger m the rejection frequencies are higher, especially for $df = 1$ and $df = 2$, which implies that the test more often flags the bootstrap distribution as non-normal. This is expected, as more data generally implies higher power to reject when normality is false, which, according to question 3, is the case here. When m is low, there are simply not enough data points to conclude non-normality. For $df = 5$, where normality is present, the rejection rate is similar to the initially set significance rate of 5%.

Question 5

The DGP from Example 3 has a single regressor and a single instrument. The value of β is left unspecified, so in the code we present an array of possible values for it (which later will be of use[1]).

For the moment, from this array we choose $\beta \approx 1.67$ for illustration purposes. We generate a sample for each possible combination of n and π. Then we estimate π with OLS and β with 2SLS[2]. Hence, we compute the T-statistic from formula (2) in the Assignment, using 1,000 MC simulations. We generate Q-Q plots to compare the distribution of the simulated T-statistic to a normal distribution (Figure 6). Alongside them we print the values of skewness and kurtosis.

When the instrument is irrelevant, the T-statistic is clearly not normal. Instead, if the instrument is strong, or even weak, the T-statistic adheres more to a normal distribution. Increasing the sample size does not change the interpretation significantly.

Question 6

We generate 4 samples of size $n = 100$ for different values of the instrument's relevance (c), choosing again $\beta \approx 1.67$ for illustration purposes. For each sample, we estimate $\hat{\pi}_{OLS}$ and $\hat{\beta}_{2SLS}$, which allows us to obtain fitted residuals $\hat{u}$ and $\hat{\varepsilon}$. Then, we run a residual bootstrap, resampling residuals from $\hat{u}$ and $\hat{\epsilon}$ for $m = n$ times, and we generate the bootstrapped T-statistic[3]. We simulate this procedure with 1,000 MC iterations and create Q-Q plots of the conditional bootstrap distribution of the T-statistic (Figure 7).

The properties observed in the original sample are also reflected in the bootstrap samples. When the instrument is irrelevant, the bootstrapped T-statistic does not follow a normal distribution and has

[1] At the end of the appendix, we show that the normality test in a bootstrapped small sample and with an irrelevant regressor interestingly gives slightly different results for different values of β.

[2] As we have only one regressor and one instrument, the model is just-identified and 2SLS simplifies to IV.

[3] During our computations, we also considered the possibility of standardizing the bootstrapped T-statistic beforehand, using its 'theoretical' standard error derived from the 2SLS estimation. In this way, we would have obtained the formula for the standardized T-statistic which is found at the end of page 3 of the Assignment. However, as noted by Cameron and Trivedi (2005), some problems may arise from the computation of the standard errors in a 2SLS. Moreover, in small samples it can give misleading results. So, in order to obtain more reliable results, we decided to not standardize the T-statistic beforehand, doing it instead in the case-specific context of Q-Q plots' generation and non-normality tests.

a similar shape to the Q-Q plots showed earlier. The points deviate significantly from the 45° line in both the left and right quantiles, indicating the presence of heavy tails in the distribution of the T-statistic. On the other hand, when the instrument is weakly, moderately or strongly relevant, the T-statistic approximates a normal distribution quite well, as expected.

Question 7

We generate a sample for each possible combination of n, π, and β[4]. Analogously to Question 6, we estimate $\hat{\pi}_{OLS}$ and $\hat{\beta}_{2SLS}$ and obtain fitted residuals. Then, we run a residual bootstrap (resampling residuals $m = n^{4/5}$ times) and we generate the bootstrapped T-statistic. We simulate this procedure using 1,000 MC iterations.

At each iteration we run a Shapiro-Wilk normality test on the bootstrapped T-statistics. We collect the results in a table with rejection frequencies. Figure 1 shows an excerpt from the table for $\beta \approx 1.67$. The full table (with more values for β) is in the Appendix (Figure 8).

Increasing the sample size has an effect on our normality diagnostics, albeit a limited one. Indeed, when the instrument is irrelevant, non-normality becomes more evident as the sample size increases, until we are practically always able to reject the null hypothesis of Gaussianity at a 5% significance level. This implies that bootstrapping produces more reliable results when the initial sample is larger.

Shapiro-Wilk Normality test: Monte Carlo simulation summary

1000 MC replications; bootstrapping with m=$n^{4/5}$

yellow cells = rejection frequency > 90%; red cells = rejection frequency > 99%.

Sample size	Instrument relevance	β	Mean p-value	Rejection frequency at 5%
100	irrelevant	1.67	0.01	97.90
	weak	1.67	0.50	4.80
	strong	1.67	0.49	4.60
200	irrelevant	1.67	0.00	99.90
	weak	1.67	0.48	5.50
	strong	1.67	0.51	5.20
400	irrelevant	1.67	0.00	100.00
	weak	1.67	0.48	5.20
	strong	1.67	0.47	6.00

Figure 1: Shapiro-Wilk Normality test.

Question 8

The purpose of Question 8 is to test if the empirical distribution of bootstrapping statistics deviates significantly from the standard normal distribution by calculating $\hat{d}_n$ (Kolmogorov-Smirnov (KS)-type distance measure). In order to calculate the critical value, we first generate the dataset under $\beta = 0$ and compute the bootstrap statistics T_n^*, which is then used for the computation of the empirical cumulative distribution $\hat{G}_n(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{T_{n:i}^* \leq x\}$. The distance measure is given by $\hat{d}_n = \sup_{x \geq 1.96} |\hat{G}_n(x) - \Phi(x)|$, where $\Phi(x)$ is the standard normal cumulative distribution function. We repeat over 1,000 MC simulations. The critical value is determined by calculating the $(1-\alpha) - th$ quintile of the null

[4]The values of β taken into consideration are contained in the array mentioned in our answer to Question 5 and defined in our code.

distribution of $\hat{d}_n$. Rejection frequencies are shown in Figure 9, with critical values in Figure 10. The table shows the rejection frequencies when $\beta \approx 1.67$.

We observe a similar behaviour in rejection frequencies as in Question 7, as shown in Figure 9. Indeed, when $\pi = 0$, the rejection rates become close to 100% and higher as the sample size increases. However, rejection rates are virtually indistinguishable from 0 for higher values of π.

Question 9

Inspired by Cameron and Trivedi (2005, pp. 376–377), we suggest to use the paired bootstrap and the wild bootstrap to address heteroskedasticity. These methods use the same original test statistic and rejection rule. The statistic is $T_n = \sqrt{n}(\hat{\beta}_n - \beta)$, where $\hat{\beta}$ is the 2SLS (or IV) estimator. Normality is rejected when T_n exceeds the critical value generated from the bootstrap distribution.

The **paired bootstrap** resamples (y_i, x_i), i.e., both the regressor and the dependent variable are resampled. This scheme allows the conditional variance of u_i to vary with regressor. The bootstrap test statistic is $T_n^* = \sqrt{n}(\hat{\beta^*}_n - \beta)$. This method does not provide an asymptotic refinement since no restriction is imposed on the expected value of the error term.

In the **wild bootstrap**, we first run the regression to get the residuals from the original model: $\hat{u}_i = x_i - \pi z_i$ and $\hat{\varepsilon}_i = y_i - \beta x_i$. Hence, we replace the OLS residuals by the folllowing residuals:

$$\hat{u}_i^* = \begin{cases} -0.618\hat{u}_i, & \text{with probability } 0.7236, \\ 1.618\hat{u}_i, & \text{with probability } 0.2764, \end{cases}, \quad \hat{\varepsilon}_i^* = \begin{cases} -0.618\hat{\varepsilon}_i, & \text{with probability } 0.7236, \\ 1.618\hat{\varepsilon}_i, & \text{with probability } 0.2764. \end{cases}$$

Then, we construct the bootstrap sample by $x_i^* = \pi z_i + u_i^*$, $y_i^* = \beta x_i^* + \varepsilon_i^*$. Results of resampling vary due to different realizations of the bootstrap residuals. The bootstrap test statistic is the same as in the paired bootstrap test statistic.

Appendix

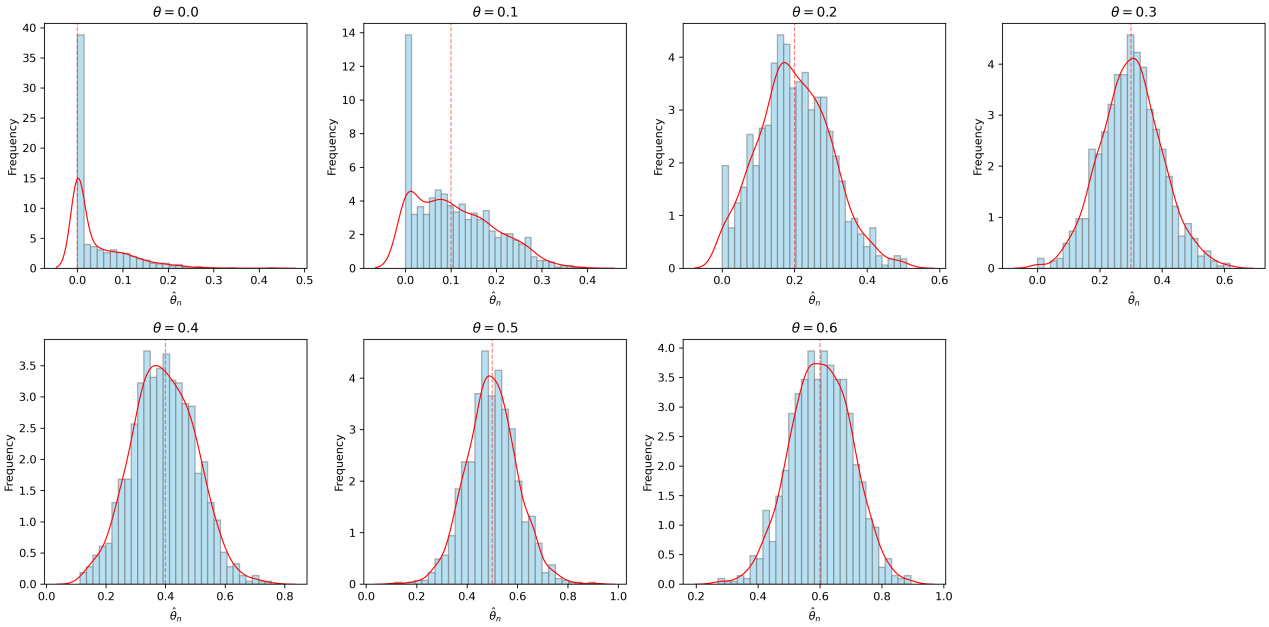

Figure 2: *Simulations of $\hat{\theta}_n$ for different values of θ (Question 1).*

	Test	Anderson-Darling	Doornik-Hansen	Jarque-Bera	Lilliefors	Shapiro-Wilk
Sample size	**Theta**					
100	**0.0**	91.70	89.20	86.50	89.20	94.90
	0.1	64.50	60.10	50.70	59.60	73.50
	0.2	29.30	26.00	20.30	24.80	37.00
	0.3	8.00	8.70	5.80	7.50	11.50
	0.4	4.40	5.00	4.00	5.90	4.90
	0.5	4.00	5.70	5.30	4.70	5.60
	0.6	4.40	4.70	3.90	4.90	4.80
400	**0.0**	95.10	95.60	94.80	93.90	97.20
	0.1	41.30	44.90	40.50	35.90	56.80
	0.2	5.90	6.70	5.60	5.70	7.90
	0.3	4.00	4.90	4.10	4.90	5.10
	0.4	3.70	6.10	5.10	5.00	4.90
	0.5	5.10	5.90	5.10	5.60	5.10
	0.6	4.60	4.30	3.90	5.70	4.40

Figure 3: *Rejection % for different normality tests.*
Yellow cells = rejection frequency > 25%; red cells = rejection frequency > 50% (Question 2).

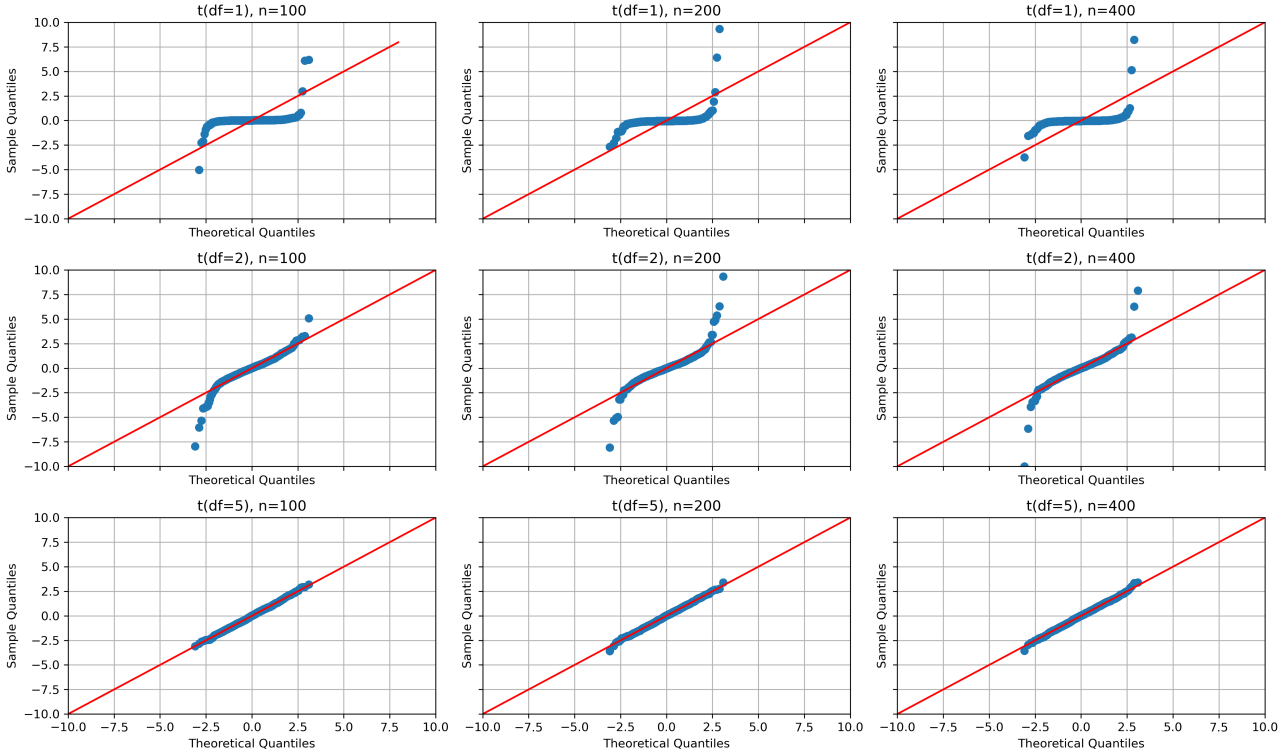

Figure 4: *Q-Q plots for different dof of the t-distribution and sample sizes (Question 3).*

				Rejection frequency at 5%
Sample size	**Degrees of freedom (v)**	**m_label**	**m**	
100	**1**	**n**	**100**	50.20
		n/2	**50**	40.70
		n^(1/2)	**10**	11.30
		n^(4/5)	**39**	33.10
	2	**n**	**100**	18.30
		n/2	**50**	14.30
		n^(1/2)	**10**	7.00
		n^(4/5)	**39**	12.10
	5	**n**	**100**	6.70
		n/2	**50**	5.60
		n^(1/2)	**10**	5.30
		n^(4/5)	**39**	5.60
200	**1**	**n**	**200**	65.50
		n/2	**100**	48.90
		n^(1/2)	**14**	17.20
		n^(4/5)	**69**	48.50
	2	**n**	**200**	23.30
		n/2	**100**	18.70
		n^(1/2)	**14**	6.80
		n^(4/5)	**69**	15.40
	5	**n**	**200**	5.20
		n/2	**100**	4.70
		n^(1/2)	**14**	5.40
		n^(4/5)	**69**	5.40
400	**1**	**n**	**400**	74.20
		n/2	**200**	64.10
		n^(1/2)	**20**	21.50
		n^(4/5)	**120**	56.20
	2	**n**	**400**	30.70
		n/2	**200**	23.40
		n^(1/2)	**20**	7.50
		n^(4/5)	**120**	16.40
	5	**n**	**400**	5.90
		n/2	**200**	5.70
		n^(1/2)	**20**	5.20
		n^(4/5)	**120**	5.20

Figure 5: *Rejection % for t-distribution with different dof, sample sizes and bootstrap sample sizes. Yellow cells = rejection frequency > 25%; red cells = rejection frequency > 50% (Question 4).*

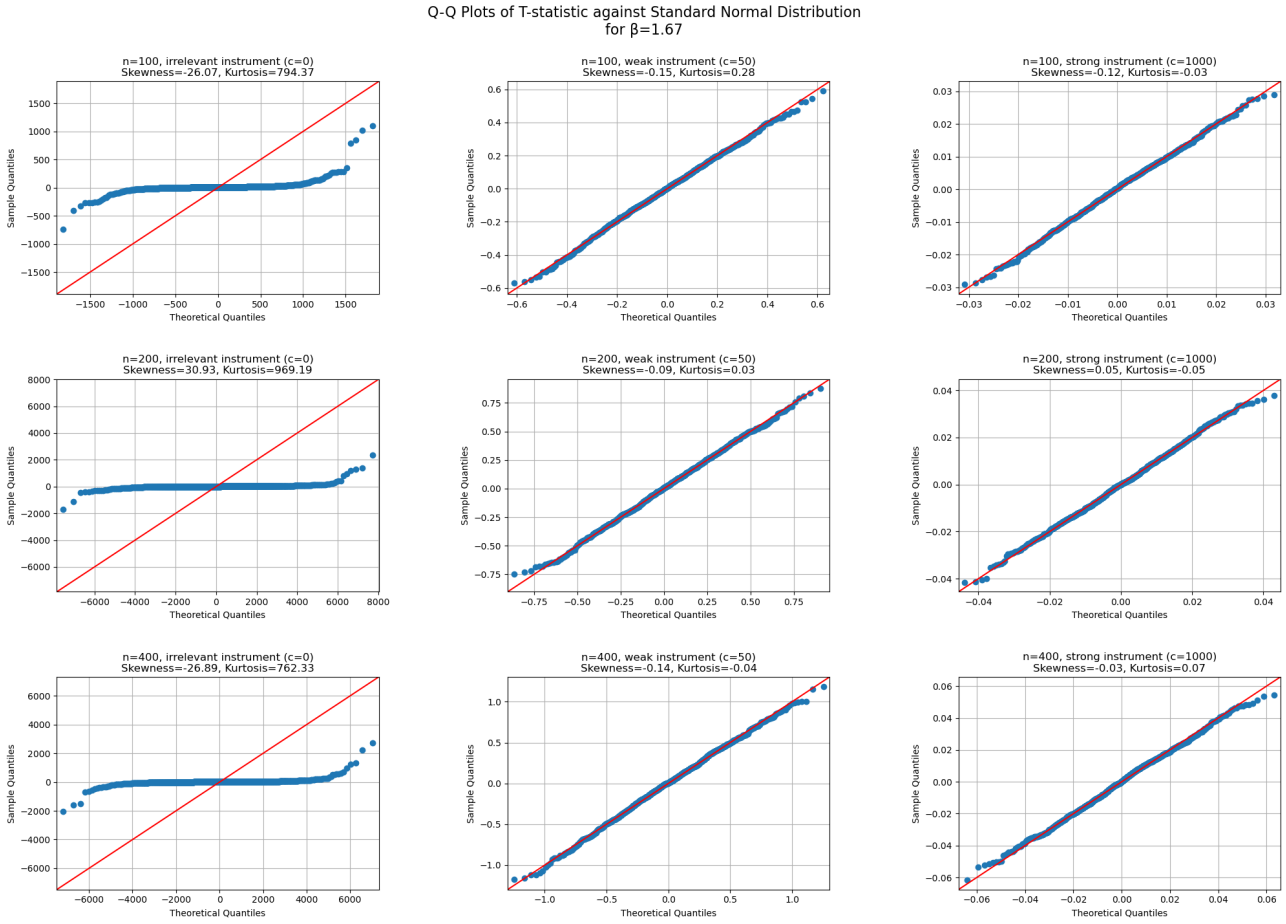

Figure 6: *Q-Q plots of T-statistic (Question 5).*

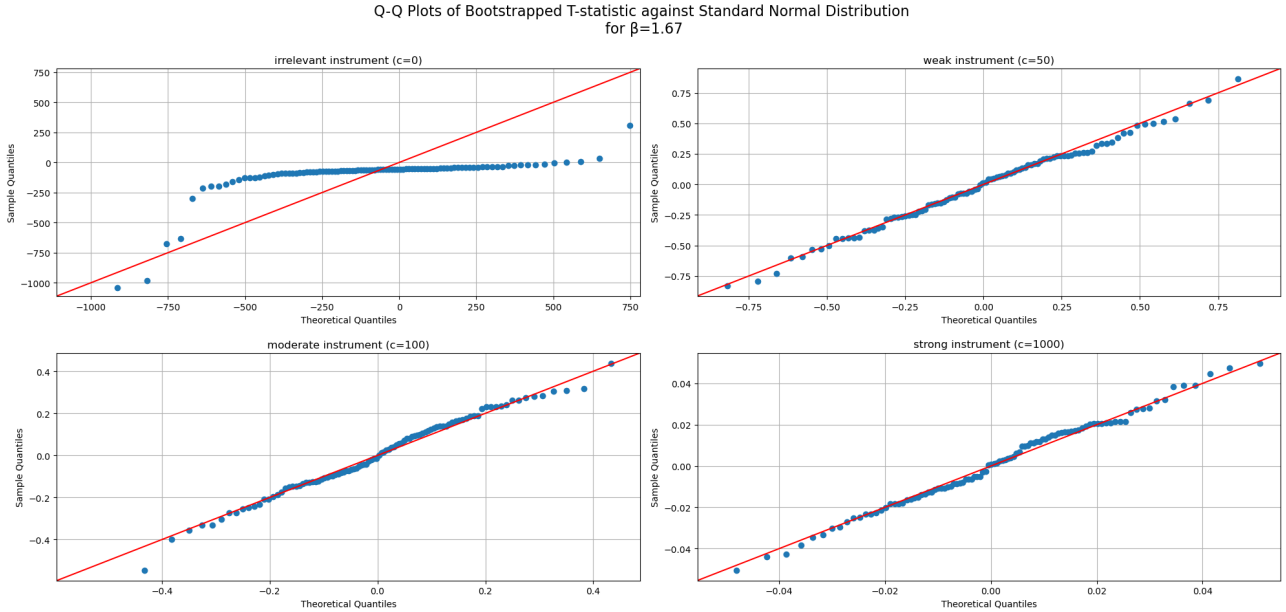

Figure 7: *Q-Q plots of Bootstrapped T-statistic (Question 6).*

Shapiro-Wilk Normality test: Monte Carlo simulation summary

1000 MC replications; bootstrapping with $m=n^{4/5}$

yellow cells = rejection frequency > 90%; red cells = rejection frequency > 99%.

Sample size	Instrument relevance	β	Mean p-value	Rejection frequency at 5%	Rejection frequency at 1%
100	irrelevant	-3.00	0.01	97.20	95.90
		-2.33	0.01	98.00	96.00
		-1.67	0.01	98.20	96.40
		-1.00	0.01	97.20	95.90
		-0.33	0.01	97.80	95.80
		0.33	0.01	97.60	96.30
		1.00	0.01	98.00	95.80
		1.67	0.01	97.90	96.30
		2.33	0.01	97.10	95.80
		3.00	0.01	98.10	96.50
	weak	-3.00	0.50	5.30	0.60
		-2.33	0.50	6.00	1.50
		-1.67	0.50	4.50	1.10
		-1.00	0.50	6.30	1.10
		-0.33	0.49	6.40	0.80
		0.33	0.52	5.50	1.50
		1.00	0.50	5.70	1.10
		1.67	0.50	4.80	1.10
		2.33	0.50	5.40	1.50
		3.00	0.52	5.40	1.30
	strong	-3.00	0.50	5.00	0.80
		-2.33	0.50	5.10	1.00
		-1.67	0.52	4.90	1.30
		-1.00	0.50	5.50	0.80
		-0.33	0.50	4.80	0.90
		0.33	0.50	5.00	0.90
		1.00	0.48	5.50	0.80
		1.67	0.49	4.60	0.70
		2.33	0.49	4.60	0.80
		3.00	0.51	3.60	0.70
	irrelevant	-3.00	0.00	99.80	99.50
		-2.33	0.00	99.30	99.20
		-1.67	0.00	99.50	99.20
		-1.00	0.00	99.40	99.30
		-0.33	0.00	99.60	99.10
		0.33	0.00	99.40	99.00
		1.00	0.00	99.50	99.10
		1.67	0.00	99.90	99.40
		2.33	0.00	99.60	99.50
		3.00	0.00	99.30	99.10
		-3.00	0.48	6.70	1.20
		-2.33	0.50	5.80	1.40
		-1.67	0.48	6.50	2.20

200	weak	**-1.00**	0.50	5.50	0.70
		-0.33	0.49	5.80	1.80
		0.33	0.49	4.40	0.90
		1.00	0.48	6.60	1.30
		1.67	0.48	5.50	1.20
		2.33	0.50	5.40	1.20
		3.00	0.50	5.80	1.20
	strong	**-3.00**	0.51	4.30	1.10
		-2.33	0.50	4.20	0.80
		-1.67	0.49	4.10	1.30
		-1.00	0.51	4.70	0.50
		-0.33	0.50	4.40	0.70
		0.33	0.51	4.00	0.60
		1.00	0.49	5.40	1.10
		1.67	0.51	5.20	1.20
		2.33	0.52	4.40	1.00
		3.00	0.49	5.20	0.90
400	irrelevant	**-3.00**	0.00	99.70	99.70
		-2.33	0.00	99.80	99.80
		-1.67	0.00	100.00	100.00
		-1.00	0.00	100.00	100.00
		-0.33	0.00	99.70	99.60
		0.33	0.00	99.80	99.80
		1.00	0.00	100.00	100.00
		1.67	0.00	100.00	99.90
		2.33	0.00	100.00	100.00
		3.00	0.00	100.00	100.00
	weak	**-3.00**	0.48	7.50	1.30
		-2.33	0.48	5.50	0.90
		-1.67	0.48	6.30	1.90
		-1.00	0.48	4.80	2.20
		-0.33	0.48	5.00	1.50
		0.33	0.48	5.40	1.70
		1.00	0.49	6.90	2.00
		1.67	0.48	5.20	1.60
		2.33	0.48	5.40	1.80
		3.00	0.49	5.50	1.40
	strong	**-3.00**	0.50	4.50	1.00
		-2.33	0.49	4.20	0.70
		-1.67	0.50	4.00	0.70
		-1.00	0.49	4.40	1.00
		-0.33	0.49	5.10	0.70
		0.33	0.50	4.30	0.70
		1.00	0.49	4.40	1.10
		1.67	0.47	6.00	0.70
		2.33	0.48	4.60	0.50
		3.00	0.51	4.70	1.30

Figure 8: *Full table with rejection frequencies (Question 7).*

Rejection Frequencies
1000 MC replications; bootstrap-based critical values.

Sample size	Instrument relevance	β	Mean d_n	Rejection frequency (%)
100	irrelevant	1.666667	0.41	93.80
	strong	1.666667	0.02	0.00
	weak	1.666667	0.02	0.00
200	irrelevant	1.666667	0.44	96.70
	strong	1.666667	0.02	0.00
	weak	1.666667	0.02	0.00
400	irrelevant	1.666667	0.45	98.30
	strong	1.666667	0.02	0.00
	weak	1.666667	0.02	0.00

Figure 9: *Full table with rejection frequencies (Question 8).*

	Sample size	Critical value
0	100	0.055036
1	200	0.038292
2	400	0.029929

Figure 10: *Critical values (Question 8).*

References

Cameron, A. Colin and Pravin K. Trivedi (May 2005). *Microeconometrics: Methods and Applications*. 1st ed. Cambridge University Press. ISBN: 978-0-521-84805-3 978-0-511-81124-1. DOI: 10.1017/CBO9780511811241. URL: https://www.cambridge.org/core/product/identifier/9780511811241/type/book (visited on 05/11/2024).