Financial Innovation

RESEARCH

Open Access

Scale elasticity and technical efficiency measures in two-stage network production processes: an application to the insurance sector

Alireza Amirteimoori[1*], Tofigh Allahviranloo[1] and Aliasghar Arabmaldar[2]

*Correspondence: alireza.amirteimoori@istinye.edu.tr

[1] Faculty of Engineering and Natural Sciences, Istinye University, Istanbul, Turkey
[2] Department of Business Administration, Faculty of Business and Economics, University of Goettingen, 37073 Göttingen, Germany

Abstract

In performance analysis with tools such as data envelopment analysis, calculations of scale properties of the frontier points are studied using both qualitative and quantitative approaches. When the production process is a bit complicated, the calculation needs to be modified. Most existing studies are focused on a single-stage production process under the constant or variable returns to scale specification. However, some processes have two-stage structures, and, in such processes, the concepts of scale elasticity and returns to scale are inextricably related to the conditions of the stages of production. Thus, an evaluation of efficiency, scale elasticity, and returns to scale is sensitive to stages. In this study, we introduced a procedure to calculate technical efficiency and scale elasticity in a two-stage parallel-series production system. Then, our proposed technical efficiency and scale elasticity programs are applied to real data on 20 insurance companies in Iran. After applying our estimations to a real-world insurance industry, we found that, (i) overall, the total inputs of insurers in the life insurance sector should be reduced by 9%. Moreover, the inputs of nonlife insurers should be reduced by 50%. The final output in the investment sector must be increased by 48%. (ii) There are inefficiencies among all insurers in the investment sector, and to improve technical efficiency, the income from investments should be increased significantly. (iii) Finally, the efficiency and elasticity characterizations of insurers are directly subject to stages.

Keywords: Data envelopment analysis, Returns to scale, Scale elasticity, Technical efficiency, Insurance companies

Introduction

In nonparametric performance analysis methods such as data envelopment analysis (DEA), several economic concepts, such as returns to scale (RTS), marginal rates of productivity and substitutions, and economies of scope, have been frequently studied. The first DEA-based work (Banker et al. 1984) and subsequent extensions that calculated the RTS characterization of frontier points in the production technology set considered a simple firm with initial inputs and final outputs, but the network structure of the

Springer Open

© The Author(s) 2024, corrected publication 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

production process was ignored. However, many production processes have a network structure that is arranged in series or parallel. In such production systems, the processes often comprise two or more subprocesses, and the outputs from one stage are used as inputs for the next stage. The two-stage production processes in the DEA setting may arise in different situations. Discussion on how to compute the scale elasticity, RTS, and marginal rates of technical substitutions in such network-structured production processes is an attractive topic in the field of DEA, attracting the attention of many scholars.

In evaluating the RTS characterizations of firms with tools such as DEA, two different approaches—qualitative and quantitative approaches—are used frequently. The former characterizes the RTS type as increasing, decreasing, or constant. However, the latter deals with the computation of scale elasticity in a quantitative form. Most studies that apply qualitative approaches or computational methods to evaluate the scale economies of frontier points focused on firms with a single-stage process under the constant or variable RTS specification.

To the best of our knowledge, there is no DEA-based research that has calculated the scale elasticity or RTS in the qualitative form of decision-making units (DMUs) in a parallel-series production process. Previous studies took the work process in the insurance sector as a black box. Thus, they did not consider intermediate products, raising questions about the validity of the analysis. In two-stage production relationships, the concepts of scale elasticity and RTS are inextricably related to the conditions of the stages of production. When there are multiple stages, the evaluation of efficiency, scale elasticity, and RTS becomes sensitive to stages and links between them. Therefore, we determine whether the efficiency and elasticity characterizations of firms are subject to stages.

In this study, the production process has a two-stage parallel-series structure in which the first stage comprises two subprocesses arranged in series. The second stage is fed by the outputs generated from the subprocesses to generate final outputs. We focus on the calculation of technical efficiency, scale elasticity, and RTS in the two-stage parallel-series production processes. To set up measures of efficiency and scale elasticity, first, we formulate a DEA program to evaluate the relative efficiency of DMUs with two-stage structures. Second, we set up the dual of this linear program to compute the scale elasticity scores of the subprocesses and stages. To demonstrate the applicability of the proposed procedure, an illustrative empirical application is provided in which the Iranian insurance data are divided into two stages (services and investment parts) and two subprocesses (life and nonlife insurance sectors). It is important to examine the efficiency and RTS characteristics of insurers across two categories and two sectors—service and investment categories and life and nonlife insurance sectors. Although the presented empirical example on Iranian insurers is illustrative, the proposed scale elasticity computation procedure can be applied to characterize the scale properties of many real-world problems whose underlying production processes have two-stage structures arranged in parallel and series.

In summary, the main contributions of this study are twofold. In the theoretical part, a procedure is proposed to evaluate the relative efficiency, scale elasticity, and RTS of DMUs with a two-stage process. Our proposed approach can be used to estimate the relative efficiency and scale elasticity of real-life sectors whose underlying production processes are two-stage. Examples of such sectors are banking, healthcare, agriculture,

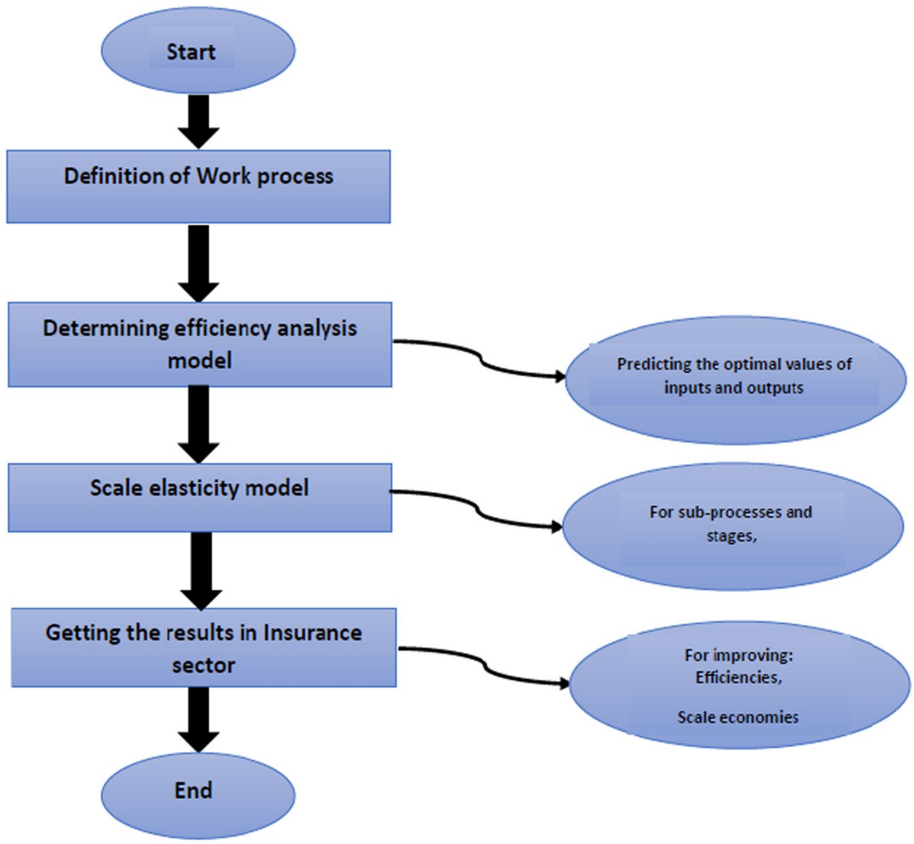

Fig. 1 Conceptual framework of the research

manufacturing, product development, and high-tech. Then, in the application part, the technical efficiency and scale elasticity in the insurance sector are examined as a two-stage process. The conceptual framework of this study is depicted in Fig. 1, highlighting both the methodological and applied approaches.

The rest of the paper is organized as follows. In Section "Literature review", we briefly review the existing works related to the subject. In Section "Measure of technical efficiency in a parallel-series production system", we first state the main problem and then propose a measure of technical efficiency in a parallel-series production system. In Section "Scale elasticity measure", we set up a procedure for computing scale elasticity scores in stages and processes. Section "An application to the insurance sector" applies the proposed scale elasticity computation procedure to data on 20 insurance companies in Iran. Finally, the study concludes in Section "Conclusions".

Literature review

DEA is a linear programming-based approach for estimating the technical efficiency of homogeneous DMUs with multiple inputs and outputs. In the last two decades, DEA has been widely applied to industrial sectors (e.g., Akbarian 2020; Dagar et al. 2021; Zhang et al. 2022; Zakari et al. 2022; Khan et al. 2022; Nourani et al. 2022; Dagar and Malik 2023; Çolak and Koy 2023; Guru et al. 2023). Traditional DEA models (the

CCR model of Charnes et al. 1978) and the BCC model of Banker et al. (1984) consider the reference technology set of a black box production unit as a single-stage procedure. To account for the network structure underlying any production process, different authors have proposed various approaches in the DEA framework to calculate efficiency and RTS. Färe and Grosskopf (2000) conducted an in-depth study of the production process to evaluate the performance of an organization and its components. For more references on network production process, see the studies by Jahanshahloo et al. (2004), Prieto and Zofio (2007), Kao and Hwang (2008), Kao (2009), Cook et al. (2010), Lewis et al. (2013), Amirteimoori (2013), Sahoo et al. (2014a, b), Sahoo et al. (2014a, b), Jelassi and Delhoumi (2021), and Kremantzis et al. (2022).

The problem with estimating RTS and scale elasticity in network production processes is one of the most frequently studied subjects in the field of DEA. Although a lot of research papers have applied qualitative approaches or computational methods to evaluate scale economies of frontier points, they focused on firms with a single-stage process (e.g., Färe et al. 1985; Banker and Thrall 1992; Førsund 1996; Sueyoshi 1999; Fukuyama 2000; Banker et al. 2004; Podinovski et al. 2009, 2016; Zelenyuk 2013; Sahoo and Tone 2013; Krivonozhko et al. 2014; Balk et al. 2015; Lee 2021; Amirteimoori et al. 2023). To the best of our knowledge, only a few studies have calculated scale elasticity and RTS in network-structured production processes (see Khaleghi et al. 2012; Patrizii 2020; Sarparast et al. 2022). Khaleghi et al. (2012) introduced an approach for estimating the nature of RTS in a two-stage process by considering the SE quantity in each of the individual stages. In this approach, the stages are arranged in series. In many real cases, the first stage in this approach is considered a black box.

The applications of parametric and nonparametric techniques in the insurance sector and other financial sectors are not rare. Eling and Luhnen (2010a) provided empirical evidence on the measurement of efficiency in the international insurance industry. Different methodologies, countries, organizational forms, and company sizes have been compared, considering both life and nonlife insurers. Kou et al. (2014) used MCDM methods to evaluate clustering algorithms for financial risk analysis. Zhao et al. (2021) investigated the evolution and determinants of the profitability of 53 Chinese insurers from 2013 to 2017. The profitability efficiency DEA model was measured, and a profit ratio change index was applied to compare the performance of Chinese insurers. Omrani et al. (2022) proposed a multiobjective network DEA model to evaluate insurance companies in an uncertain environment. Li et al. (2022) employed an integrated cluster-detection, optimization, and interpretation approach to study financial data. Smętek et al. (2022) surveyed and critically evaluated the literature on the use of advanced DEA-based methods to assess the financial effectiveness of insurance entities. Banker et al. (2022) compared Indian and Iranian insurance companies and calculated the managerial ability of the insurers of the two countries. More references on the applications of the parametric and nonparametric techniques in the insurance sector are the studies by Yang (2006), Cummins et al. (2010), Cummins and Rubio-Misas (2021), Kou et al. (2019), Kou et al. (2021a, b), Kou et al. (2021a, b), and Frederick et al. (2022). For a complete overview of frontier efficiency measurement in the insurance sector, see the study by Eling and Luhnen (2010b).

Measure of technical efficiency in a parallel-series production system

Consider the two-stage parallel-series production process portrayed in Fig. 2, in this process, the first stage comprises two different subprocesses arranged in parallel. These two subprocesses operate separately under the supervision of the whole process. Outputs generated by these two subprocesses are used as inputs for the second stage. Thus, the second stage is fed by two intermediate products to generate the final outputs. For each $MU_j, j \in \{1, 2, \ldots, n\}$. $x_j^{(k)} = \left(x_{1j}^{(k)}, x_{2j}^{(k)}, \ldots, x_{\mathrm{m}_k j}^{(k)}\right)^T$ and $z_j^{(k)} = \left(z_{1j}^{(k)}, z_{2j}^{(k)}, \ldots, z_{\mathrm{D}_k j}^{(k)}\right)^T$ are inputs and outputs of the subprocess $k, (k = 1, 2)$, respectively. $z_j^{(1)}$ and $z_j^{(2)}$ are used as inputs for the second stage to produce the final outputs$\mathrm{y}_j = \left(\mathrm{y}_{1j}, \mathrm{y}_{2j}, \ldots, \mathrm{y}_{sj}\right)^T$.

The intermediate measures $z^{(1)}$ and $z^{(2)}$ are dual-role variables in the sense that they are the outputs of the first stage and are used as inputs for the second stage. To evaluate the technical efficiency of DMU_o (DMU under evaluation) in output orientation, we set up the following variable RTS-DEA model:

$$
\begin{aligned}
& E_o^* = \mathrm{Max}\varepsilon \\
& \text{s.t.} \\
& \sum_{j=1}^{n} \lambda_j^{(1)} x_{ij}^{(1)} \leq x_{io}^{(1)}, i = 1, \ldots, m_1, \\
& \sum_{j=1}^{n} \lambda_j^{(1)} z_{dj}^{(1)} \geq \widetilde{z}_{do}^{(1)}, d = 1, \ldots, D_1, \\
& \sum_{j=1}^{n} \lambda_j^{(1)} = 1, \\
& \sum_{j=1}^{n} \lambda_j^{(2)} x_{ij}^{(2)} \leq x_{io}^{(2)}, i = 1, \ldots, m_2, \\
& \sum_{j=1}^{n} \lambda_j^{(2)} z_{dj}^{(2)} \geq \widetilde{z}_{do}^{(2)}, d = 1, \ldots, D_2, \\
& \sum_{j=1}^{n} \lambda_j^{(2)} = 1, \\
& \sum_{j=1}^{n} \mu_j y_{rj} \geq \varepsilon y_{ro}, r = 1, \ldots, s, \\
& \sum_{j=1}^{n} \mu_j z_{dj}^{(1)} \leq \widetilde{z}_{do}^{(1)}, d = 1, \ldots, D_1, \\
& \sum_{j=1}^{n} \mu_j z_{dj}^{(2)} \leq \widetilde{z}_{do}^{(2)}, d = 1, \ldots, D_2, \\
& \sum_{j=1}^{n} \mu_j = 1, \\
& \lambda_j, \mu_j \geq 0, \quad \forall j.
\end{aligned}
\tag{1}
$$

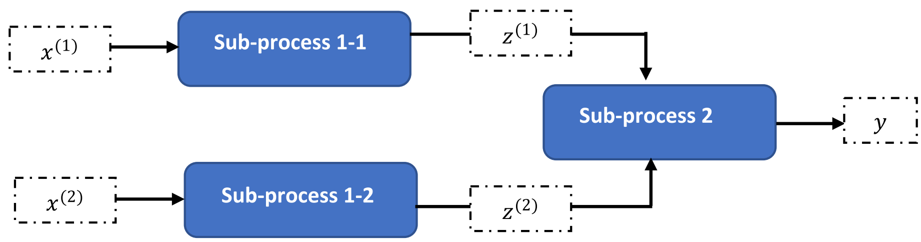

Fig. 2 Two-stage parallel-series network process

The first three constraints in Model (1) are related to Subprocess 1; the second three constraints are related to Subprocess 2; and the last four constraints are included in Stage 2. In Model (1), DMU_o is efficient if and only if $E_o^* = 1$. The dual formulation of Model (1) can be written as follows:

$$
\begin{aligned}
& E_o^* = Max \sum_{i=1}^{m_1} v_i^{(1)} x_{io}^{(1)} + \sum_{i=1}^{m_2} v_i^{(2)} x_{io}^{(2)} + w_o^{(1)} + w_o^{(2)} + u_0 \\
& \text{s.t.} \\
& \sum_{i=1}^{m_1} v_i^{(1)} x_{ij}^{(1)} - \sum_{d=1}^{D_1} w_d^{(1)} z_{dj}^{(1)} + w_o^{(1)} \geq 0, \forall j, \\
& \sum_{i=1}^{m_2} v_i^{(2)} x_{ij}^{(2)} - \sum_{d=1}^{D_2} w_d^{(2)} z_{dj}^{(2)} + w_o^{(2)} \geq 0, \forall j, \\
& \sum_{d=1}^{D_1} \mu_d^{(1)} z_{dj}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{dj}^{(2)} - \sum_{r=1}^{s} u_r y_{rj} + u_0 \geq 0, \forall j, \\
& \sum_{r=1}^{s} u_r y_{ro} = 1, \\
& \sum_{d=1}^{D_1} w_d^{(1)} - \sum_{d=1}^{D_1} \mu_d^{(1)} = 0, \\
& \sum_{d=1}^{D_2} w_d^{(2)} - \sum_{d=1}^{D_2} \mu_d^{(2)} = 0, \\
& v_i^{(1)}, v_i^{(2)}, w_d^{(1)}, w_d^{(2)}, \mu_d^{(1)}, \mu_d^{(2)} \geq 0.
\end{aligned}
\tag{2}
$$

Let $\left(v^{(1)}, v^{(2)}, w^{(1)}, w^{(2)}, \mu^{(1)}, \mu^{(2)}, w_o^{(1)}, w_o^{(2)}, u, u_o\right)$ be the optimal solution to Model (2). The technical efficiencies of the stages and subprocesses of DMU_o are defined as follows:

$$
TE_o^{(1,1)} = \frac{\sum_{i=1}^{m_1} v_i^{(1)} x_{io}^{(1)} + w_o^{(1)}}{\sum_{d=1}^{D_1} w_d^{(1)} z_{do}^{(1)}}, \quad TE_o^{(1,2)} = \frac{\sum_{i=1}^{m_2} v_i^{(2)} x_{io}^{(2)} + w_o^{(2)}}{\sum_{d=1}^{D_2} w_d^{(2)} z_{do}^{(2)}},
$$

$$
TE_o^{(2)} = \frac{\sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)} + u_0}{\sum_{r=1}^{s} u_r y_{ro}},
$$

Following Chen et al. (2009), we compute a posteriori TE_o^N of DMU_o for the whole process as a positive linear combination of its component efficiencies, i.e.,

$$TE_o^N = \rho_1 E_o^{(1,1)} + \rho_2 TE_o^{(1,2)} + \rho_3 TE_o^2, \tag{3}$$

where the weights ρ_1, ρ_2, and ρ_3 are defined as follows:

$$\begin{aligned}
\rho_1 &= \frac{\sum_{d=1}^{D_1} w_d^{(1)} z_{do}^{(1)}}{1 + \sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)}} \\
\rho_2 &= \frac{\sum_{d=1}^{D_1} w_d^{(1)} z_{do}^{(1)}}{1 + \sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)}}, \text{ and} \\
\rho_3 &= \frac{\sum_{r=1}^{s} u_r y_{ro}}{1 + \sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)}}
\end{aligned}$$

The weights ρ_1, ρ_2, and ρ_3 are determined endogenously by our dual evaluation program (Model (2)), and each unit is different.

By substituting the values of $\rho_{1,}\rho_2$, and ρ_3 into Eq. (3) we derive the following:

$$\begin{aligned}
TE_o^N &= \rho_1 E_o^{(1,1)} + \rho_2 TE_o^{(1,2)} + \rho_3 TE_o^2 \\
&= \frac{\sum_{d=1}^{D_1} w_d^{(1)} z_{do}^{(1)}}{1 + \sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)}} \times \frac{\sum_{i=1}^{m_1} v_i^{(1)} x_{io}^{(1)} + w_o^{(1)}}{\sum_{d=1}^{D_1} w_d^{(1)} z_{do}^{(1)}} \\
&\quad + \frac{\sum_{d=1}^{D_1} w_d^{(1)} z_{do}^{(1)}}{1 + \sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)}} \times \frac{\sum_{i=1}^{m_2} v_i^{(2)} x_{io}^{(2)} + w_o^{(2)}}{\sum_{d=1}^{D_2} w_d^{(2)} z_{do}^{(2)}} \\
&\quad + \frac{\sum_{r=1}^{s} u_r y_{ro}}{1 + \sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)}} \times \frac{\sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)} + u_0}{\sum_{r=1}^{s} u_r y_{ro}} \\
&= \frac{\sum_{i=1}^{m_1} v_i^{(1)} x_{io}^{(1)} + w_o^{(1)} + \sum_{i=1}^{m_2} v_i^{(2)} x_{io}^{(2)} + w_o^{(2)} + \sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)} + u_0}{1 + \sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)}} \\
&= \frac{E_o^* + \sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)}}{1 + \sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)}}
\end{aligned} \tag{4}$$

In this case, $TE_o^N = 1$ if and only if $E_o^* = 1$.

Scale elasticity measure

Consider the frontier point $\left(x_o^{(1)}, x_o^{(2)}, z_o^{(1)}, z_o^{(2)}, y_o\right)$ in production technology set and focus on stage $k(k = 1\text{or}2)$. For any α_k close to one, we define $\overline{\beta}_k(\alpha_k)$ as the response function of the changes in intermediate products $z_o^{(k)}$ to changes in initial inputs $x_o^{(k)}$. Clearly, when the initial inputs proportionally change by α_k, it influences intermediate measures $z_o^{(k)}$ and final outputs y_o. Therefore, we first need to calculate the effect of changes in initial inputs $x_o^{(k)}$ on intermediate measures $z_o^{(k)}$. We also need to calculate the effect of changes in intermediate measures $z_o^{(k)}$ on final outputs. In this sense, we first need to calculate the response function $\overline{\beta}_k(\alpha_k)$ by solving the following program:

$$
\begin{aligned}
&\overline{\beta}_k(\alpha_k) = Max\beta_k \\
&\text{s.t.} \\
&\sum_{j=1}^{n} \lambda_j x_{ij}^{(k)} \le \alpha_k x_{io}^{(k)}, i = 1, \ldots, m_k, \\
&\sum_{j=1}^{n} \lambda_j z_{dj}^{(k)} \ge \beta_k z_{do}^{(k)}, d = 1, \ldots, D_k, \\
&\sum_{j=1}^{n} \lambda_j = 1, \\
&\lambda_j \ge 0.
\end{aligned}
\tag{5}
$$

Note that α_k is a user-defined value close to 1, reflecting the proportional changes in initial inputs. Consider the dual formulation of Model (5) as follows:

$$
\begin{aligned}
\overline{\beta}_k(\alpha_k) =& Min\alpha_k \sum_{i=1}^{m_k} v_i^{(k)} x_{io}^{(k)} + u_0^{(k)} \\
&\text{s.t.} \\
&\sum_{i=1}^{m_k} v_i^{(k)} x_{io}^{(k)} - \sum_{d=1}^{D_1} w_d^{(k)} z_{dj}^{(k)} + u_0^{(k)} \ge 0, \forall j \\
&\sum_{d=1}^{D_k} w_d^{(k)} z_{do}^{(k)} = 1, \\
&v_i^{(k)}, w_d^{(k)} \ge 0, \forall i, d, r, u_0 \text{free}
\end{aligned}
\tag{6}
$$

To compute the scale elasticity of the *k-th* process of the first stage, we consider the following transformation function of firm *o*:

$$
\psi\left(\alpha_k x_o^{(k)}, \beta_k z_o^{(k)}\right) = \alpha_k \sum_{i=1}^{m_k} v_i^{(k)} x_{io}^{(k)} - \beta_k \sum_{d=1}^{D_k} w_d^{(k)} z_{do}^{(k)} + u_0^{(k)} = 0, \tag{7}
$$

Assume that the transformation function (7) is differentiable, and differentiating (7) related to the input-scaling factor α_k, we derive the following:

$$
\frac{\partial\psi\left(\alpha_k z_o^{(k)}, \beta_k z_o^{(k)}\right)}{\partial\alpha_k} = \sum_{r=1}^{s} \frac{\partial\psi\left(\alpha_k x_o^{(k)}, \beta_k z_o^{(k)}\right)}{\partial\left(\alpha_k x_o^{(k)}\right)} x_o^{(k)} - \sum_{d=1}^{D_k} \frac{\partial\psi\left(\alpha_k x_o^{(k)}, \beta_k z_o^{(k)}\right)}{\partial\left(\beta_k z_o^{(k)}\right)} z_{do}^{(k)} \frac{\partial\overline{\beta}_k(\alpha_k)}{\partial\alpha_k} = 0
$$

$$
\begin{aligned}
&\Rightarrow \frac{\partial\overline{\beta}_k(\alpha_k)}{\partial\alpha_k} = \frac{\sum_{r=1}^{s} \frac{\partial\psi\left(\alpha_k x_o^{(k)}, \beta_k z_o^{(k)}\right)}{\partial\left(\alpha_k x_o^{(k)}\right)} x_o^{(k)}}{\sum_{d=1}^{D_k} \frac{\partial\psi\left(\alpha_k x_o^{(k)}, \beta_k z_o^{(k)}\right)}{\partial\left(\beta_k z_o^{(k)}\right)} z_{do}^{(k)}} = \frac{\sum_{i=1}^{m_k} v_i^{(k)} x_{io}^{(k)}}{\sum_{d=1}^{D_k} w_d^{(k)} z_{do}^{(k)}} \\
&= \frac{\frac{\overline{\beta}_k(\alpha_k) - u_0^{(k)}}{\alpha_k}}{1} = \frac{\overline{\beta}_k(\alpha_k) - u_0^{(k)}}{\alpha_k}.
\end{aligned}
\tag{8}
$$

The scale elasticity of the *k-th* process of the first stage of firm *o* is now defined as the ratio of its marginal productivity to its average productivity as follows:

$$\varepsilon_o^{(1,k)} = \frac{\partial\overline{\beta}_k(\alpha_k)}{\partial\alpha_k}\frac{\alpha_k}{\overline{\beta}_k(\alpha_k)} = \frac{\overline{\beta}_k(\alpha_k) - u_0^{(k)}}{\alpha_k}\frac{\alpha_k}{\overline{\beta}_k(\alpha_k)} = 1 - \frac{u_0^{(k)}}{\overline{\beta}_k(\alpha_k)} \tag{9}$$

Due to the piecewise linearity of the production frontier in the DEA production set, the efficient frontier in DEA is not differentiable at the extreme points. Therefore, we set up the following two programs to determine the right-hand (left-hand) scale elasticity $\varepsilon_{1,k}^{+}\left(\varepsilon_{1,k}^{-}\right)$ of the first stage of DMU_o:

$$\begin{aligned}
&\left(u_0^{(-,k)}\right)\left(u_0^{(+,k)}\right) = Max(Min)u_0^{(k)}\\
&\text{s.t.}\\
&\alpha_k\sum_{i=1}^{m_k} v_i^{(k)}x_{io}^{(k)} + u_0^{(k)} = \overline{\beta}_k(\alpha_k),\\
&\sum_{i=1}^{m_k} v_i^{(k)}x_{io}^{(k)} - \sum_{d=1}^{D_1} w_d^{(k)}z_{dj}^{(k)} + u_0^{(k)} \geq 0, \forall j\\
&\sum_{d=1}^{D_k} w_d^{(k)}z_{do}^{(k)} = 1,\\
&v_i^{(k)}, w_d^{(k)} \geq 0, \forall i, d, r, u_0^{(k)}\ free
\end{aligned} \tag{10}$$

Using the optimal values of Model (10), the right- and left-hand side scale elasticities of firm *o* can be defined as $\varepsilon_{1,k}^{+} = 1 - \frac{u_0^{(+,k)}}{\overline{\beta}_k(\alpha_k)}$ and $\varepsilon_{1,k}^{-} = 1 - \frac{u_0^{(-,k)}}{\overline{\beta}_k(\alpha_k)}$, respectively.

Theorem 1. *The RTS characterization of the k-th subprocess of the first stage of firm o is as follows*:

(a) *decreasing RTS* (DRS) (*i.e.*, $\varepsilon_{1,k}^{-} < 1$) *if* $u_0^{(-,k)} > 0$
(b) *constant RTS* (CRS) (*i.e.*, $\varepsilon_{1,k}^{-} \leq 1 \leq \varepsilon_{1,k}^{+}$) *if* $u_0^{(+,k)} \leq 0 \leq u_0^{(-,k)}$
(c) *increasing RTS* (IRS) (*i.e.*,$\varepsilon_{1,k}^{+} > 1$) *if*$u_0^{(+,k)} < 0$.

Proof. Assume that$v^{(k)*}$, $w^{(k)*}$, and $u_0^{(k)*}$ are the optimal solutions of Model (6) for efficient point$(x_o^{(k)}, z_o^{(k)})$. Then, $\sum_{d=1}^{D_k} w_d^{(k)}z_{do}^{(k)} = 1$. Moreover, $\sum_{i=1}^{m_k} v_i^{(k)}x_{io}^{(k)} - \sum_{d=1}^{D_k} w_d^{(k)}z_{dj}^{(k)} + u_0^{(k)} = 0$ is a supporting surface of the production technology set in Stage 1 and the corresponding tangent at $\left(\overline{\beta}_k(\alpha_k), \alpha_k\right)$ is presented in Model (8). The slope of this tangent is $\varepsilon(\gamma) = \frac{\alpha_k - u_0^{(k)*}}{\overline{\beta}_k(\alpha_k)}$.
Therefore, the right- and left-hand slopes are $\varepsilon^{+}(\pi) = \frac{\overline{\beta}_k(\alpha_k) - u_0^{(+,k)}}{\overline{\beta}_k(\alpha_k)}$ and $\varepsilon^{-}(\pi) = \frac{\overline{\beta}_k(\alpha_k) - u_0^{(-,k)}}{\overline{\beta}_k(\alpha_k)}$, respectively.

(a) This indicates that $\varepsilon^{+}(\pi) < 1$ if $u_0^{(+,k)} > 0$, and as $u_0^{(-,k)} > u_0^{(+,k)}$, $\varepsilon^{-}(\pi) > 1$, indicating an IRS at this point.

(b) Further, if $u_0^{(+,k)} < 0$, $\varepsilon^+(\pi) > 1$, and if $u_0^{(-,k)} > 0$, $\varepsilon^-(\pi) < 1$. Thus, $\varepsilon^-(\pi) \le 1 \le \varepsilon^+(\pi)$, indicating a CRS at this point.
(c) Finally, if $u_0^{(+,k)} > 0$, $\varepsilon^+(\pi) < 1$, and as $u_0^{(+,k)} < u_0^{(-,k)}$, $\varepsilon^-(\pi) < 1$, indicating a DRS at this point.

Now, we are ready to calculate the scale elasticity of the second stage of the process. To do this, we assume that $\beta = Min\{\beta_1, \beta_2\}$ and calculate the response function $\overline{\pi}(\beta)$ as follows:

$$\begin{aligned}
&\overline{\pi}(\beta) = Max\pi \\
&\text{s.t.} \\
&\sum_{j=1}^{n} \lambda_j y_{rj} \ge \pi y_{ro}, i = 1, \ldots, m_k, \\
&\sum_{j=1}^{n} \lambda_j z_{dj}^{(1)} \le \beta z_{do}^{(1)}, d = 1, \ldots, D_1, \\
&\sum_{j=1}^{n} \lambda_j z_{dj}^{(2)} \le \beta z_{do}^{(2)}, d = 1, \ldots, D_2, \\
&\sum_{j=1}^{n} \lambda_j = 1, \\
&\lambda_j, \pi \ge 0.
\end{aligned} \tag{11}$$

β is the minimum value of two response values obtained from the two subprocesses of the first stage that reflects the proportional changes in intermediate products $z_o^{(1)}$ and $z_o^{(2)}$. The dual formulation of Model (11) is as follows:

$$\begin{aligned}
&\overline{\pi}(\beta) = \min \beta \sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \beta \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)} + \mu_0 \\
&\text{s.t.} \\
&\sum_{d=1}^{D_1} \mu_d^{(1)} z_{dj}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{dj}^{(2)} - \sum_{r=1}^{s} u_r y_{rj} + \mu_0 \ge 0, \forall j, \\
&\sum_{r=1}^{s} u_r y_{ro} = 1, \\
&u_r, \mu_d^{(1)}, \mu_d^{(2)} \ge 0, \forall i, d, \mu_0 \text{free}.
\end{aligned} \tag{12}$$

Consider the following transformation function of firm *o*:

$$\Phi\left(\beta z_o^{(1)}, \beta z_o^{(2)}, \pi y_o\right) = \beta \sum_{d=1}^{D_1} \mu_d^{(1)} z_{do}^{(1)} + \beta \sum_{d=1}^{D_2} \mu_d^{(2)} z_{do}^{(2)} - \pi \sum_{r=1}^{s} u_r y_{ro} + \mu_0 = 0, \tag{13}$$

Assume that the transformation function (13) is differentiable. On differentiation of (13) related to the output-scaling factor β, we derive the following:

$$\frac{\partial\Phi\left(\beta z_o^{(1)},\beta z_o^{(2)},\pi y_o\right)}{\partial\beta}=\sum_{d=1}^{D_1}\frac{\partial\Phi\left(\beta z_o^{(1)},\beta z_o^{(2)},\pi y_o\right)}{\partial\left(\beta z_{do}^{(1)}\right)}z_{do}^{(1)}+\sum_{d=1}^{D_2}\frac{\partial\Phi\left(\beta z_o^{(1)},\beta z_o^{(2)},\pi y_o\right)}{\partial\left(\beta z_{do}^{(2)}\right)}z_{do}^{(2)}$$
$$-\frac{\partial\Phi\left(\beta z_o^{(1)},\beta z_o^{(2)},\pi y_o\right)}{\partial\left(\pi y_{ro}\right)}y_{ro}\frac{\partial\overline{\pi}(\beta)}{\partial\beta}=0.$$

Thus, we derive:

$$\Phi\frac{\partial\overline{\pi}(\beta)}{\partial\beta}=\frac{\sum_{d=1}^{D_1}\frac{\partial\Phi\left(\beta z_o^{(1)},\beta z_o^{(2)},\pi y_o\right)}{\partial\left(\beta z_{do}^{(1)}\right)}z_{do}^{(1)}+\sum_{d=1}^{D_2}\frac{\partial\Phi\left(\beta z_o^{(1)},\beta z_o^{(2)},\pi y_o\right)}{\partial\left(\beta z_{do}^{(2)}\right)}z_{do}^{(2)}}{\frac{\partial\Phi\left(\beta z_o^{(1)},\beta z_o^{(2)},\pi y_o\right)}{\partial\left(\pi y_{ro}\right)}y_{ro}}$$
$$=\frac{\sum_{d=1}^{D_1}\mu_d^{(1)}z_{do}^{(1)}+\sum_{d=1}^{D_2}\mu_d^{(2)}z_{do}^{(2)}}{\sum_{r=1}^{s}u_r y_{ro}}=\frac{\frac{\overline{\pi}(\beta)-\mu_0}{\beta}}{1}=\frac{\overline{\pi}(\beta)-\mu_0}{\beta} \tag{14}$$

The scale elasticity of the second stage of firm *o* is now calculated as follows:

$$\varepsilon_o^{(2)}=\frac{\partial\overline{\pi}(\beta)}{\partial\beta}\frac{\beta}{\overline{\pi}(\beta)}=\frac{\overline{\pi}(\beta)-\mu_0}{\beta}\frac{\beta}{\overline{\pi}(\beta)}=\frac{\overline{\pi}(\beta)-\mu_0}{\overline{\pi}(\beta)}=1-\frac{\mu_0}{\overline{\pi}(\beta)} \tag{15}$$

Again, as the efficient frontier in DEA is not differentiable at the extreme efficient points, we set up the following two models to determine the right- and left-hand scale elasticities (ε_2^{+}andε_2^{-}) for the second stage of firm *o*:

$$\begin{aligned}\left(\mu_0^{(-)}\right)\left(\mu_0^{(+)}\right)=&\max(\min)\mu_0\\ &\text{s.t.}\\ &\beta\sum_{d=1}^{D_1}\mu_d^{(1)}z_{do}^{(1)}+\beta\sum_{d=1}^{D_2}\mu_d^{(2)}z_{do}^{(2)}+\mu_0=\overline{\pi}(\beta)\\ &\sum_{d=1}^{D_1}\mu_d^{(1)}z_{dj}^{(1)}+\sum_{d=1}^{D_2}\mu_d^{(2)}z_{dj}^{(2)}-\sum_{r=1}^{s}u_r y_{rj}+\mu_0\geq 0,\forall j,\\ &\sum_{r=1}^{s}u_r y_{ro}=1,\\ &u_r,\mu_d^{(1)},\mu_d^{(2)}\geq 0,\forall i,d,\mu_0\text{isfree}\end{aligned} \tag{16}$$

Using the optimal values of Model (16), the right- and left-hand sides scale elasticities of firm *o* can be calculated as $\varepsilon_2^{-}=1-\frac{\mu_0^{(-)}}{\overline{\pi}(\beta)}$ and $\varepsilon_2^{+}=1-\frac{\mu_0^{(+)}}{\overline{\pi}(\beta)}$.

Theorem 2. *The RTS characterization of the second stage of firm o is as follows*:

(*a*) *DRS, i.e.,* $\varepsilon_{1-k}^{+}<1$ *if* $\mu_0^{(+k)}>0$
(*b*) *CRS, i.e.,* $\varepsilon_{1-k}^{+}\leq 1\leq\varepsilon_{1-k}^{-}$ *if* $\mu_0^{(-k)}\leq 0\leq\mu_0^{(+k)}$
(*c*) *IRS, i.e.,* $\varepsilon_{1-k}^{-}<1$ *if* $\mu_0^{(-k)}>0$.

Proof. Let u^*, $\mu^{(1)*}$, $\mu^{(2)*}$, and μ_0^* be the optimal solutions of the linear program (12) for efficient point $(z_o^{(1)}, z_o^{(2)}, y_o)$. Then, $\sum_{r=1}^{s} u_r y_{ro} = 1$. Moreover, $\sum_{d=1}^{D_1} \mu_d^{(1)} z_{dj}^{(1)} + \sum_{d=1}^{D_2} \mu_d^{(2)} z_{dj}^{(2)} - \sum_{r=1}^{s} u_r y_{rj} + \mu_0 = 0$ is a supporting surface of the production set of the second stage, and the corresponding tangent at $(\beta, \overline{\pi}(\beta))$ is given by Model (14). The slope of this tangent is $\varepsilon = \frac{\overline{\pi}(\beta) - \mu_0^*}{\overline{\pi}(\beta)}$.
Hence, we can compute the right- and left-hand slopes and derive the following:

$$\varepsilon^+ = \frac{\overline{\pi}(\beta) - \mu_0^{(+)*}}{\overline{\pi}(\beta)} \varepsilon^- = \frac{\overline{\pi}(\beta) - \mu_0^{(-)*}}{\overline{\pi}(\beta)}$$

(a) This indicates that $\varepsilon^- > 1$ if $\mu_0^{(-)*} < 0$, and as $\mu_0^{(+)*} < \mu_0^{(-)*}$, $\varepsilon^+ > 1$, indicating an IRS at this point.
(b) Further, if $\mu_0^{(+)*} < 0$, $\varepsilon^+ > 1$, and if $\mu_0^{(-)*} > 0$, $\varepsilon^- < 1$. Thus, $\varepsilon^- \leq 1 \leq \varepsilon^+$, indicating a CRS at this point.
(c) Finally, if $\mu_0^{(+)*} > 0$, $\varepsilon^+ < 1$, and as $\mu_0^{(+)*} < \mu_0^{(-)*}$, $\varepsilon^- < 1$, indicating a DRS at this point.

The scale elasticity of the whole system $\varepsilon_o^{(W)}$ is defined as the marginal productivity of the whole system to the average productivity as follows:

$$\varepsilon_o^{(W)} = \frac{\partial \overline{\pi}\left(\overline{\beta}(\alpha_k)\right)}{\partial \alpha_k} \times \frac{\alpha_k}{\overline{\pi}\left(\overline{\beta}(\alpha_k)\right)}. \tag{17}$$

Theorem 3. *Assume that* $\varepsilon_o^{(1-k)} : k = 1, 2$ *and* $\varepsilon_o^{(2)}$ *are the scale elasticities of Subprocesses* 1 *and* 2 *of the first stage and the scale elasticity of Stage* 2, *respectively. Then, the scale elasticity of the whole system is calculated as follows*:

$$\varepsilon_o^{(W)} = \frac{1}{2}[\varepsilon_o^{(1-1)} + \varepsilon_o^{(1-2)}]\varepsilon_o^{(2)}.$$

Proof. Using Eq. (17) to compute the scale elasticity of the whole process, we derive the following:

$$\begin{aligned}
\varepsilon_o^{(W)} &= \frac{\partial \overline{\pi}\left(\overline{\beta}(\alpha_k)\right)}{\partial \alpha_k} \times \frac{\alpha_k}{\overline{\pi}\left(\overline{\beta}(\alpha_k)\right)} \\
&= \frac{\partial \overline{\pi}\left(\overline{\beta}(\alpha_k)\right)}{\partial \overline{\beta}(\alpha_k)} \times \frac{\partial \overline{\beta}(\alpha_k)}{\partial \alpha_k} \times \frac{\alpha_k}{\overline{\pi}\left(\overline{\beta}(\alpha_k)\right)} \\
&= \left(\frac{\overline{\pi}(\beta) - \mu_0}{\beta}\right) \times \left(\frac{\overline{\beta}(\alpha_k) - u_0^{(k)}}{\alpha_k}\right) \times \frac{\alpha_k}{\overline{\beta}(\alpha_k)} \times \frac{\overline{\beta}(\alpha_k)}{\overline{\pi}\left(\overline{\beta}(\alpha_k)\right)} = \varepsilon_o^{(2)} \times \varepsilon_o^{(1-k)}
\end{aligned}$$

This implies that $\varepsilon_o^{(W)} = \frac{1}{2}[\varepsilon_o^{(1,1)} + \varepsilon_o^{(1,2)}]\varepsilon_o^{(2)}$, completing the proof.

We now demonstrate the scale elasticity computation approach with a small-scale example consisting of three DMUs. The first stage uses one initial input for each of Subprocesses 1 and 2. Each of these two subprocesses generates its own output, which is used as input for the second stage. Thus, the second stage is fed by two inputs—the

Table 1 The data set for simple example

Banks	$x^{(1)}$	$x^{(2)}$	$z^{(1)}$	$z^{(2)}$	y	Efficiency
A	1	2	1	1	2	1
B	2	3	4	4	3	1
C	4	5	6	5	3	1

Table 2 The scale elasticity scores of 20 insurers

DMU	Stage 1						Stage 2			Whole system		
	Subprocess 1			Subprocess 1								
	ε^-	ε^+	RTS	ε^-	ε^+	RTS	ε^-	ε^+	RTS			
A	3	INF	I	6	INF	I	0.1667	INF	C	0.75015	INF	C
B	0.5	1.5	C	0.3750	2.2500	C	0	0.4444	D	0	0.83325	D
C	0	0.6667	D	0	0.5	D	0	0	D	0	0	D

I, IRS; C, CRS; and D, DRS

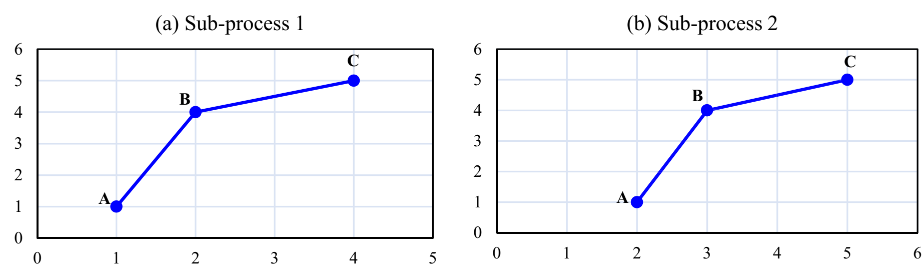

Fig. 3 Production set for simple example

single output of each of Subprocesses 1 and 2 from Stages 1 and 2 generates a single output as a final output. The dataset is listed in the first five columns of Table 1.

All DMUs are efficient in stages and the whole system. For example, in DMU_A, the left-hand scale elasticity of Subprocess 1 in the first stage is 3, and the right-hand scale elasticity is $+\infty$. However, the left-hand scale elasticity of Subprocess 2 in the first stage is 6, and the right-hand scale elasticity is $+\infty$. Referring to Theorem 1, we find that the RTS characterization of both subprocesses is increasing. Regarding the RTS characterization of the second stage, the 8th and 9th columns of Table 2 indicate that DMU_A has a CRS. Finally, the whole system has a CRS.

The production sets of the two subprocesses are depicted in Fig. 3. In both subprocesses, DMU_A has an IRS; the RTS of DMU_B is constant; and that of DMU_C is decreasing. These are confirmed by the results presented in Table 2.

An application to the insurance sector

To demonstrate the applicability of our proposed approach, we apply it to a real case of insurance companies. Optimizing the relative efficiency of insurance companies is one of the ways to maintain a competitive advantage. Increasing market share, maximizing revenue, and reducing costs are the main goals of insurance companies, and they compete to optimize their performance. Previous research on the applications of DEA in the insurance sector considered the reference technology set of a black box production unit as a single- or two-stage procedure.

Data and variables

The current section focuses on estimating the relative efficiency and scale elasticity of the insurance market in Iran. Each insurance company operates in two sectors—life and nonlife insurance services. As these two sectors do business separately under the supervision of a central decision-maker, for illustration purposes, we considered each sector as a separate subprocess. A systemic view of the work process of the insurance companies is depicted in Fig. 4. The first stage comprises two subprocesses—life and nonlife insurance sectors. The revenues obtained from these two sectors are used to invest in the second stage, and the income from investments is considered the final output of the process. Thus, the work process of insurance companies is considered a two-stage procedure. Regarding the selection of inputs and outputs, efficiency analysis in the insurance sector requires the specification of inputs and outputs of insurers. Eling and Luhnen (2010a, b) considered three types of inputs—labor, business services, and capital. Leverty and Grace (2010) considered three alternative approaches for choosing outputs—the financial intermediation approach, the user cost approach, and the value-added approach. The value-added approach introduced by Berger et al. (1987) and Berger and Humphrey (1992) consider activities that contribute significant value-added as outputs, which are assessed using operating cost allocations.

In some studies, net premiums are used as valued-added, while in others, incurred benefits and changes in reserves are used as output proxies. In this section, we intend to evaluate the relative efficiencies and scale elasticities of companies, with a specific focus on revenue generation. We are interested in determining the main output indicators that contribute to generating revenue. Therefore, we selected two outputs—net premium income and other incomes. We also considered two inputs—operating expenses (expenses that are related to the insurance business) and claims incurred (claims paid in the period). Claim incurred is an undesirable output as it is the outcome of transforming input to output, but we are interested in reducing the level of claim incurred. One way to accommodate an undesirable output is to treat it as input (to be reduced). Thus, we considered the claim incurred as an input. This study includes 20 insurance companies in Iran from 2019 to 2020 (a description of the variables is presented in Table 3).

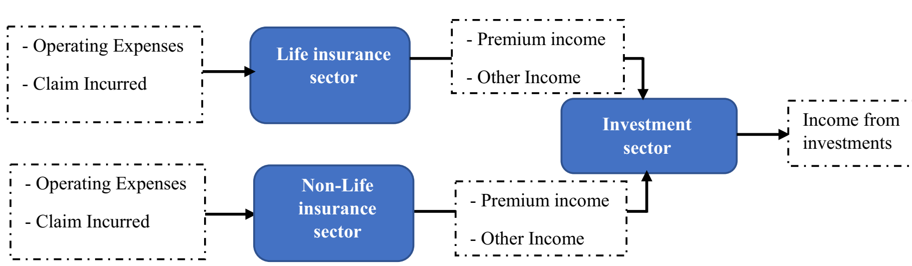

Fig. 4 A systemic view of the work process in insurance sector

Table 3 Description of three different variables

Description	Indicators	Type of variable
Life Insurance sector	Operating expenses, claims incurred	Initial input variables
	Net premium income, other income	Output variables (Intermediate)
Non-life insurance sector	Operating expenses, claims incurred	Initial input variables
	Net premium income, other income	Output variables (Intermediate)
Investment sector	Income from investments	Final output variable

Table 4 A summary statistic of the input and output data

Index	Life insurance sector				Non-life insurance sector				Investment sector
	Inputs		Outputs		Inputs		Outputs		Final outputs
	OE	CI	NPI	OI	OE	CI	NPI	OI	IFI
MIN	153.17	105.77	101.84	100.98	134.00	130.00	103.00	104.00	785.08
MAX	2955.62	3841.59	1858.20	4090.31	7791.00	7149.00	9451.00	2790.00	2515.77
MEAN	1137.51	1067.70	700.99	1393.45	978.40	1984.10	2420.95	613.25	1455.08
STD	697.36	957.18	583.19	905.14	1717.90	2245.14	2786.39	817.26	599.14
MEDIAN	987.06	800.75	506.40	1255.31	382.00	960.00	1205.00	210.50	1221.39

An important point is that the inputs and outputs of the life and nonlife insurance sectors are the same; thus, we can combine these two sectors in a single stage. However, here, these two parts are considered as a black box, so we cannot analyze the performance of the two sectors. This is why we split the first stage into two different subprocesses.

The summary statistic of the input–output data is listed in Table 4.

Results

We first applied our proposed efficiency Model (1) to this dataset. The last column of Table 5 presents the statistical description of the efficiency results. It indicates that out of 20 insurers, 5 were efficient. As scale elasticity is defined and calculated only for frontier units, for inefficient insurers, this calculation is performed only for their input- or output-oriented projections of the efficient frontier. To calculate the scale elasticity of the insurers, we consider their output-oriented projection points. We used Model (1) to project inefficient insurers onto the efficient frontier. The statistical description of the optimal values of the input–output variables is presented in Table 5. Overall, the total inputs of the insurers in the life insurance sector must be reduced by 9%. Moreover, the inputs of nonlife insurers must be reduced by 50%. The final output in the investment sector must be increased by 48%.

We first calculated the scale elasticity of the subprocesses of the first stage along with the scale elasticity of the second stage. The results on both the lower and upper bounds of the scale elasticities for all 20 insurers are listed in the first six columns of Table 6. As an example, the results on Insurer 1, which is efficient, are interpreted as follows: In the first subprocess, its lower- and upper-bound scale elasticity scores are less than 1, i.e., $< \varepsilon^- < \varepsilon^+ < 1$, implying that it has a DRS. In Subprocess 2, its lower-bound scale

Table 5 A summary statistic of the optimal values of the inputs and outputs

Index	Life insurance sector				Non-life insurance sector				Investment sector	Total efficiency score
	Inputs		Outputs		Inputs		Outputs		Final outputs	
	OE	CI	NPI	OI	OE	CI	NPI	OI	IFI	
MIN	153.17	105.77	195.68	1060.88	134.00	130.00	103.00	104.00	1385.44	1.00
MAX	2955.62	3841.59	1812.54	4090.31	769.00	2118.00	2924.00	458.00	2515.77	2.57
MEAN	1037.06	968.02	884.23	2107.67	430.11	1121.62	1510.91	320.27	2157.53	1.71
STD	680.69	961.04	477.36	853.73	243.99	746.57	1038.52	122.41	246.31	0.64
MEDIAN	925.26	727.49	996.11	1922.05	382.00	938.32	1251.43	327.09	2236.86	1.77

Table 6 The scale elasticity scores of 20 insurers

Banks	Stage 1						Stage 2		
	Subprocess 1			Subprocess 1					
	ε^-	ε^+	RTS	ε^-	ε^+	RTS	ε^-	ε^+	RTS
1	0.0000	0.6416	D	0.9805	1.0132	C	0.0000	0.7884	D
2	0.7904	0.8468	D	1.0248	1.0437	I	0.2955	1.4470	C
3	0.8110	0.8757	D	1.0029	1.0331	I	0.3277	2.0077	D
4	0.3578	6.8267	D	1.0045	1.0523	I	4.2035	INF	I
5	0.5311	9.5384	D	0.9542	1.0311	C	0.2561	INF	C
6	0.7068	1.4917	C	0.9858	1.0096	C	0.5627	1.6752	C
7	0.7885	0.8181	D	1.1129	1.0238	I	0.7830	0.8154	D
8	0.4130	INF	C	1.0024	1.0284	I	0.6122	5.4097	C
9	0.4130	INF	C	1.0031	1.0350	I	15.9632	INF	I
10	0.0000	1.5670	C	0.9901	1.0146	C	0.0085	0.9728	D
11	0.4130	INF	C	0.7309	1.5031	C	3.4646	INF	I
12	0.0143	1.4329	C	0.9590	1.0279	C	9.5962	11.1025	I
13	0.6612	0.7149	D	0.9924	1.0052	C	0.5828	1.6440	C
14	− 0.5991	− 0.2047	D	0.0000	1.0177	C	3.4654	4.4475	I
15	0.1996	0.8000	D	0.9955	1.0266	C	0.7598	0.7956	D
16	0.6693	1.3071	C	0.9922	1.0053	C	0.6113	1.6000	C
17	0.6836	1.5093	C	0.8713	1.0875	C	1.5565	INF	I
18	0.5628	0.6646	D	0.9350	1.0992	C	0.1042	0.9415	D
19	0.5737	0.6780	D	0.4042	INF	C	0.0750	INF	C
20	0.5456	0.6534	D	0.8887	1.0757	C	0.6172	INF	C

I, IRS; C, CRS; and D, DRS

elasticity score is 0.9805, and its upper-bound scale elasticity score is 1.0132, which is greater than 1, i.e., $\varepsilon^- < 1, \varepsilon^+ > 1$, implying that it has a CRS. For the second stage, both scale elasticities are less than 1, having a DRS. Similar to Insurer 1, for efficient Insurers 15 and 18, the first subprocess of the first stage has a DRS; the second subprocess has a CRS; and the second stage has a DRS.

For efficient Insurer 2, the lower and upper bounds elasticity scores in the first subprocess are both less than 1 (i.e., $\varepsilon^- < \varepsilon^+ < 1$), indicating a DRS. However, the lower and upper bounds elasticity scores in the second subprocess are both greater than 1 (i.e., $1 < \varepsilon^- < \varepsilon^+$), indicating an IRS. In the second stage, the lower-bound scale elasticity score is 0.2955 and its upper-bound scale elasticity score is 1.4470, indicating a CRS.

For efficient Insurers 5, 13, 19, and 20, the first subprocess has a DRS, and the second subprocess and the second stage have a CRS.

We now calculate the scale elasticity of the whole process. The results on both the lower and upper bounds of the scale elasticities are presented in Table 7. Here, Insurer 1 has a DRS in Subprocess 1, a CRS in Subprocess 2, and a DRS in Stage 2. Its lower- and upper-bound scale elasticity scores are less than 1, indicating a DRS. However, for Insurer 2, its lower-bound scale elasticity score is less than 1 and its upper-bound scale elasticity score is greater than 1, indicating a CRS. In total, 5 insurers (1, 7, 15, and 18) have a DRS; 10 insurers (2, 3, 5, 6, 8, 10, 13, 16, 19, and 20) have a CRS; and six insurers (4, 9, 11, 12, 14, and 17) have an IRS.

Table 7 The scale elasticity scores of the whole system

Banks	ε^{-}	ε^{+}	RTS
1	0.0000	0.6523	D
2	0.2682	1.3678	C
3	0.2972	1.9161	C
4	2.8632	INF	I
5	0.1902	INF	C
6	0.4762	2.0950	C
7	0.7444	0.7509	D
8	0.4332	INF	C
9	11.3027	INF	I
10	0.0042	1.2556	C
11	1.9816	INF	I
12	4.6700	13.6605	I
13	0.4819	1.4139	C
14	1.0381	2.7183	I
15	0.4540	0.7267	D
16	0.5079	1.8500	C
17	1.2101	INF	I
18	0.0780	0.8303	D
19	0.0367	INF	C
20	0.4426	INF	C

I, IRS; C, CRS; and D, DRS

Table 8 The RTS types of insurers across the stages

Stages	RTS type		
	DRS	CRS	IRS
Life insurance sector	0	8	12
Non-life insurance sector	6	14	0
Investment part	5	8	7

At the end of this application, we analyzed the RTS type across various subprocesses and stages. The results are displayed in Table 8. Out of 20 insurers, in the life insurance sector, 8 insurers have a CRS, and the other 12 have a DRS. Moreover, in the nonlife insurance sector, 6 insurers have an IRS, and 14 have a CRS. Finally, regarding investment, 5 insurers have an IRS; 8 have a CRS; and 7 have a DRS.

Conclusions

In this study, we used the DEA approach to estimate the relative efficiency score and RTS types of firms. We focused on the technical efficiency, scale elasticity, and RTS of a two-stage parallel-series production system. The results reveal that there are some significant differences in the results about the scale elasticity and the RTS types of firms in the stages. These differences are due to the existence of differences in the types of services and operational environments of the stages and substages. Based on the results obtained from the scale properties of firms and their stages, chief managers can decide to expand or limit their operations in subsequent operational periods. This study demonstrates how to change outputs with respect to changing inputs to preserve efficiency in subsequent periods.

In the theoretical analysis, our proposed scale elasticity model is applied to sample data on 20 insurance companies in Iran. The important empirical findings in our real application in the insurance industry are as follows: (i) Input consumption in the nonlife insurance sector is very high, and to improve technical efficiency, insurers are expected to reduce their input consumption significantly (approximately 50%). (ii) The insurers have performed poorly in the investment sector, and to improve technical efficiency, the income from investments should be increased by 48%. (iii) The technical efficiency and elasticity characterizations of insurance companies are directly subject to the scale property of the stages.

The empirical results on scale elasticities revealed that the technical efficiency and scale elasticity characterizations of insurers are significantly subject to the stages of the process. An important point to note is that in all retrospective performance analysis models, such as DEA and benchmarking techniques, econometrician suggests the source of inefficiencies to decision-makers. This inefficiency may be due to the excessive consumption of inputs or the production of products below expectation. Decision-makers have to make decisions to prevent excessive consumption of inputs and simultaneously increase the level of production. Moreover, by characterizing the nature of the RTS of insurers, we are providing decision-makers with insight into how outputs will quantitatively change when inputs proportionally change. This insight would help decision-makers in expanding or limiting their activities.

There are some potential research issues for future research:

- A future study can extend our proposed deterministic procedure to a case in which there is uncertainty and variability in the data.
- Our proposed procedure can be extended to analyze the marginal rates of technical substitutions of firms.
- We assumed that the stages of a firm use stage-specific inputs and outputs. A future study can propose a procedure to calculate the technical efficiency and scale elasticity of a firm with contextual or explanatory variables.

Abbreviations

DEA	Data envelopment analysis
DMU	Decision-making unit
RTS	Returns to scale
CRS	Constant returns to scale
DRS	Decreasing returns to scale
IRS	Increasing returns to scale

Acknowledgements
The authors would like to thank the Editor-in-Chief, the Associate Editor, and the anonymous reviewers for their valuable feedback on the previous version of this manuscript.

Author contributions
In this research, AA and TA wrote the body of the manuscript. AA wrote the codes and prepared the first draft of the paper. AA finalized the manuscript.

Funding
There is no research funding for this research.

Availability of data and materials
All data used in this paper are available per request.

Declarations

Competing interests
On behalf of my co-authors, I declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 22 January 2023 Accepted: 6 December 2023
Published: 28 January 2024

References

Akbarian D (2020) Overall profit Malmquist productivity index under data uncertainty. Financ Innov 6:6
Amirteimoori A (2013) A DEA two-stage decision processes with shared resources. CEJOR 21:141–151
Amirteimoori A, Sahoo B, Mehdizadeh S (2023) Data envelopment analysis for scale elasticity measurement in the stochastic case: with an application to Indian banking. Financ Innov 9(1):1–36
Balk BM, Färe R, Karagiannis G (2015) On directional scale elasticities. J Prod Anal 43(1):99–104
Banker RD, Thrall RM (1992) Estimation of returns to scale using data envelopment analysis. Eur J Oper Res 62:74–84
Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30(9):1078–1092
Banker RD, Cooper WW, Seiford LM, Thrall RM, Zhu J (2004) Returns to scale in different DEA models. Eur J Oper Res 154:345–362
Banker RD, Amirteimoori A, Sinha RP (2022) An integrated data envelopment analysis and generalized additive model for assessing managerial ability with application to the insurance industry. Decis Anal J 4:100115
Berger AN, Humphrey D (1992) Measurement and efficiency issues in commercial general insurance industry in southern and western asia banking. In: ZviGriliches (ed) Output measurement in the service sectors. University of Chicago Press, Chicago, pp 245–300
Berger AN, Hanweck G, Humphrey D (1987) Competitive viability in banking: Scale, scope and product mix economies. J Monet Econ 20(3):501–520
Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision-making units. Eur J Oper Res 2:429–444
Chen Y, Cook WD, Li N, Zhu J (2009) Additive efficiency decomposition in two-stage DEA. Eur J Oper Res 196(3):1170–1176
Çolak AB, Koy A (2023) The role of technical indicators in the intraday prediction of stock markets: artificial neural network models for Borsa Istanbul. Sci Iran. https://doi.org/10.24200/sci.2023.58490.5752
Cook WD, Liang L, Zhu J (2010) Measuring performance of two-stage network structures by DEA: A review and future perspective. Omega 38(6):423–430
Cummins JD, Rubio-Misas M (2021) Country factor behavior for integration improvement of European life insurance markets. Econ Anal Policy 72:186–202
Cummins JD, Weiss MA, Xie X, Zi H (2010) Economies of scope in financial services: A DEA efficiency analysis of the US insurance industry. J Bank Finance 34(7):1525–1539
Dagar V, Malik S (2023) Nexus between macroeconomic uncertainty, oil prices, and exports: evidence from quantile-on-quantile regression approach. Environ Sci Pollut Res 30(16):48363–48374
Dagar V, Khan MK, Alvarado R, Usman M, Zakari A, Rehman A, Murshed M, Tillaguango B (2021) Variations in technical efficiency of farmers with distinct land size across agro-climatic zones: evidence from India. J Clean Prod 315:128109
Eling M, Luhnen M (2010a) Efficiency in the international insurance industry: a cross-country comparison. J Bank Finance 34(7):1497–1509
Eling M, Luhnen M (2010b) Frontier efficiency methodologies to measure performance in the insurance industry: overview systematization, and recent developments. Geneva Pap Risk Insur Issues Pract 35(2):217–265
Färe R, Grosskopf S, Lovell CAK (1985) The measurement of efficiency of production. Bosten MA
Färe R, Grosskopf S (2000) Network DEA. Socioecon Plann Sci 34:35–49
Førsund FR (1996) On the calculation of the scale elasticity in DEA models. J Prod Anal 7(2–3):283–302
Frederick JD, Fung DWH, Yang CC, Yeh JH (2022) Individual health insurance reforms in the US: expanding interstate markets, Medicare for all, or Medicaid for all? Eur J Oper Res 297(2):753–765
Fukuyama H (2000) Returns to scale and scale elasticity in data envelopment analysis. Eur J Oper Res 125(1):93–112
Guru S, Verma S, Baheti P, Dagar V (2023) Assessing the feasibility of hyperlocal delivery model as an effective distribution channel. Manag Decis 61(6):1634–1655
Jahanshahloo G, Amirteimoori A, Kordrostami S (2004) Measuring the multi-component efficiency with shared inputs and outputs in data envelopment analysis. Appl Math Comput 155(1):283–293
Jelassi MM, Delhoumi E (2021) What explains the technical efficiency of banks in Tunisia? Evidence from a two-stage data envelopment analysis. Financ Innov 7:64
Kao C (2009) Efficiency decomposition in network data envelopment analysis: a relational model. Eur J Oper Res 192(3):949–962
Kao C, Hwang SN (2008) Efficiency decomposition in two-stage data envelopment analysis: an application to non-life insurance companies in Taiwan. Eur J Oper Res 185(1):418–429
Khaleghi M, Jahanshahloo G, Zohrehbandian M, Lotfi FH (2012) Returns to scale and scale elasticity in two-stage DEA. Math Comput Appl 17:193–202
Khan I, Zakari A, Dagar V, Singh S (2022) World energy trilemma and transformative energy developments as determinants of economic growth amid environmental sustainability. Energy Econom 108:105884
Kou G, Peng Y, Wang G (2014) Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Inf Sci 275:1–12

Kou G, Chao X, Peng Y, Alsaadi FE, Herrera-Viedma E (2019) Machine learning methods for systemic risk analysis in financial sectors. Technol Econ Dev Econ 25(5):716–742
Kou G, Olgu Akdeniz Ö, Dinçer H, Yüksel S (2021a) Fintech investments in European banks: a hybrid IT2 fuzzy multidimensional decision-making approach. Financ Innov 7(1):39
Kou G, Xu Y, Peng Y, Shen F, Chen Y, Chang K, Kou S (2021b) Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decis Support Syst 140:113429
Kremantzis MD, Beullens P, Kyrgiakos LS, Klein J (2022) Measurement and evaluation of multi-function parallel network hierarchical DEA systems. Socioecon Plann Sci 84:101428
Krivonozhko V, Førsund FR, Lychev AV (2014) Measurement of returns to scale using non-radial DEA models. Eur J Oper Res 232(3):664–670
Lee HS (2021) Efficiency decomposition of the network DEA in variable returns to scale: an additive dissection in losses. Omega 100:102212
Leverty JT, Grace MF (2010) The robustness of output measures in property-liability insurance efficiency studies. J Bank Finance 34(7):1510–1524
Lewis HF, Mallikarjun S, Sexton TR (2013) Unoriented two-stage DEA: the case of oscillating intermediate products. Eur J Oper Res 229:529–539
Li T, Kou G, Peng Y, Yu PS (2022) An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Trans Cybern 52(12):13848–13861
Nourani M, Kweh QL, Lu WM et al (2022) Operational and investment efficiency of investment trust companies: do foreign firms outperform domestic firms? Financ Innov 8:79
Omrani H, Emrouznejad A, Shamsi M, Fahimi P (2022) Evaluation of insurance companies considering uncertainty: a multi-objective network data envelopment analysis model with negative data and undesirable outputs. Socio-Econ Plann Sci 82(Part B):101306
Patrizii V (2020) On network two stages variable returns to scale Dea models. Omega 97:102084. https://doi.org/10.1016/j.omega.2019.06.010
Podinovski VV, Førsund FR, Krivonozhko VE (2009) A simple derivation of scale elasticity in data envelopment analysis. Eur J Oper Res 197(1):149–153
Podinovski V, Chambers RG, Atici KB, Deineko ID (2016) Marginal values and returns to scale for nonparametric production frontiers. Oper Res 64(1):236–250
Prieto AM, Zofio JL (2007) Network DEA efficiency in input–output models: with an application to OECD countries. Eur J Oper Res 178:292–304
Sahoo BK, Tone K (2013) Non-parametric measurement of economies of scale and scope in non-competitive environment with price uncertainty. Omega 41:97–111
Sahoo BK, Zhu J, Tone K (2014a) Decomposing efficiency and returns to scale in two-stage network systems. In: Cook WD, Zhu J (eds) Data envelopment analysis: a handbook of modeling internal structure and network (chapter 7). Springer, New York, pp 137–164
Sahoo BK, Zhu J, Tone K, Klemen BM (2014b) Decomposing technical efficiency and scale elasticity in two-stage network DEA. Eur J Oper Res 233(3):584–594
Sarparast M, Lotfi FH, Amirteimoori A (2022) Investigating the sustainability of return to scale classification in a two-stage network based on DEA models. Discrete Dyn Nat Soc. https://doi.org/10.1155/2022/8951103
Smętek K, Zawadzka D, Strzelecka A (2022) Examples of the use of Data Envelopment Analysis (DEA) to assess the financial effectiveness of insurance companies. Procedia Comput Sci 207:3924–3930
Sueyoshi T (1999) DEA duality on returns to scale (RTS) in production and cost analyses: an occurrence of multiple solutions and differences between production-based and cost-based RTS estimates. Manage Sci 45(11):1593–1608
Yang Z (2006) A two-stage DEA model to evaluate the overall performance of Canadian life and health insurance companies. Math Comput Model 43(7–8):910–919
Zakari A, Khan I, Tan D, Alvarado R, Dagar V (2022) Energy efficiency and sustainable development goals (SDGs). Energy 239(Part E):122365
Zelenyuk V (2013) A scale elasticity measures for directional distance function and its dual: theory and DEA estimation. Eur J Oper Res 228:592–600
Zhang C, Khan I, Dagar V, Saeed A, Zafar MW (2022) Environmental impact of information and communication technology: Unveiling the role of education in developing countries. Technol Forecast Soc Chang 178:121570
Zhao T, Pei R, Pan J (2021) The evolution and determinants of Chinese property insurance companies' profitability: a DEA-based perspective. J Manag Sci Eng 6(4):449–466

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.