frontiers
in Applied Mathematics and Statistics

Shapley Value Confidence Intervals for Attributing Variance Explained

Daniel Fryer[1], Inga Strümke[2]* and Hien Nguyen[3]

[1]School of Mathematics and Physics, The University of Queensland, St Lucia, QLD, Australia, [2]Simula Research Laboratory, Oslo, Norway, [3]Department of Mathematics and Statistics, La Trobe University, Melbourne, VIC, Australia

The coefficient of determination, the R^2, is often used to measure the variance explained by an affine combination of multiple explanatory covariates. An attribution of this explanatory contribution to each of the individual covariates is often sought in order to draw inference regarding the importance of each covariate with respect to the response phenomenon. A recent method for ascertaining such an attribution is via the game theoretic Shapley value decomposition of the coefficient of determination. Such a decomposition has the desirable efficiency, monotonicity, and equal treatment properties. Under a weak assumption that the joint distribution is pseudo-elliptical, we obtain the asymptotic normality of the Shapley values. We then utilize this result in order to construct confidence intervals and hypothesis tests for Shapley values. Monte Carlo studies regarding our results are provided. We found that our asymptotic confidence intervals required less computational time to competing bootstrap methods and are able to exhibit improved coverage, especially on small samples. In an expository application to Australian real estate price modeling, we employ Shapley value confidence intervals to identify significant differences between the explanatory contributions of covariates, between models, which otherwise share approximately the same R^2 value. These different models are based on real estate data from the same periods in 2019 and 2020, the latter covering the early stages of the arrival of the novel coronavirus, COVID-19.

Keywords: variable importance ranking, SHapley Additive exPlanations, R square, variance explained, linear regression, asymptotic distribution, model explanations, explainable machine learning

1. INTRODUCTION

When conducting statistical estimation and computation, the assumption of randomness of data necessitates that we address not only the problem of point estimation, but also variability quantification. In Ref. 1, variability for the coefficient of determination Shapley values were quantified via the use of bootstrap confidence intervals (CIs). Combined with the usual computational intensiveness of bootstrap resampling (see, e.g., Refs. 2 and 3; Ch. 12)), the combinatory nature of the computation of **Eq. 5** (notice that $|\mathcal{P}| = d!$) compounds the time complexity of such a method, which is already of order $\mathcal{O}(2^d)$. In this article, we seek to provide an asymptotic method for computing CIs for the Shapley values.

Our approach uses the joint asymptotic normality result of the elements in a correlation matrix, under an elliptical assumption, via [4], combined with asymptotic normality results concerning the determinants of a correlation matrix, of Refs. 5 and 6. Using these results, we derive the asymptotic joint distribution for the R^2 ($\mathcal{Z}_n$) Shapley values, which allows us to construct CIs for each of the

values and their contrasts. We assess the finite sample properties of our constructions via a comprehensive Monte Carlo study and demonstrate the use of our CIs via applications to real estate price data.

The remainder of the article proceeds as follows. In **Section 3**, we present our main results regarding the asymptotic distribution of the coefficient of determination Shapley values, and their CI constructions. In **Section 4**, we present a comprehensive Monte Carlo study of our CI construction method. In **Section 5**, we demonstrate how our results can be applied to real estate price data. Conclusions are lastly drawn in **Section 6**. All data and scripts to reproduce the analyses are available at https://github.com/ex2o/shapley_confidence.

2. BACKGROUND

2.1. The Coefficient of Determination

The multiple linear regression model (MLR) is among the most commonly applied tools for statistics inference; see Refs. 7 and 8; Part I) for thorough introductions to MLR models. In the usual MLR setting, one observes an independent and identically distributed (IID) sample of data pairs $\mathbf{Z}_i^\top = (Y_i, \mathbf{X}_i^\top) \in \mathbb{R}^{d+1}$, where $i \in [n] = \{1, \ldots, n\}$, and $d, n \in \mathbb{N}$. The MLR model is then defined via the linear relationship

$$(Y_i | \mathbf{X}_i = \mathbf{x}_i) = \beta_0 + \sum_{j=1}^d \beta_j X_{ij},$$

where $\beta^\top = (\beta_0, \ldots, \beta_n) \in \mathbb{R}^{d+1}$, and $\mathbf{X}_i^\top = (X_{i1}, \ldots, X_{id})$. We shall also write $\mathbf{Z}_i^\top = (Z_{i0}, Z_{i1}, \ldots, Z_{id})$, when it is convenient to do so. Here, $(\cdot)^\top$ is the transposition operator.

The usual nomenclature is to call the Y_i and $\mathbf{X}_i$ elements of each pair, the response (or dependent) variable and the explanatory (or covariate) vector, respectively. Here, the jth element of $\mathbf{X}_i$: X_{ij} ($j \in [d]$), is referred to as the jth explanatory variable (or the jth covariate). We may put the covariate and response pairs into a dataset $\mathcal{Z}_n = \{\mathbf{Z}_i\}_{i=1}^n$.

Let $R_{jk}(\mathcal{Z}_n)$ denote the sample correlation coefficient

$$R_{jk}(\mathcal{Z}_n) = \frac{\sum_{i=1}^n (Z_{ij} - \overline{Z}_j)(Z_{ik} - \overline{Z}_k)}{\sqrt{\sum_{i=1}^n (Z_{ij} - \overline{Z}_j)^2} \sqrt{\sum_{i=1}^n (Z_{ik} - \overline{Z}_k)^2}}, \tag{1}$$

for each $j, k \in \{0\} \cup [d]$. Here $\overline{Z}_j = n^{-1} \sum_{i=1}^n Z_{ij}$ is the sample mean of variable j. Write $\mathcal{U} \subseteq \{0\} \cup [d]$ be a nonempty subset, where $\mathcal{U} = \{u_1, \ldots, u_{|\mathcal{U}|}\}$, where $|\mathcal{U}|$ is the cardinality of $\mathcal{U}$. We refer to the matrix of correlations between the variables in $\mathcal{U}$ as

$$\mathbf{C}_n(\mathcal{U}) = \begin{bmatrix} 1 & R_{u_1 u_2}(\mathcal{Z}_n) & \cdots & R_{u_1 u_{|\mathcal{U}|}}(\mathcal{Z}_n) \\ R_{u_2 u_1}(\mathcal{Z}_n) & 1 & \cdots & R_{u_2 u_{|\mathcal{U}|}}(\mathcal{Z}_n) \\ \vdots & \vdots & \ddots & \vdots \\ R_{u_{|\mathcal{U}|} u_1}(\mathcal{Z}_n) & R_{u_{|\mathcal{U}|} u_2}(\mathcal{Z}_n) & \cdots & 1 \end{bmatrix}. \tag{2}$$

A common inferential task is to determine the degree to which the response can be explained by the covariate vector, in totality. The usual device for addressing this question is via the coefficient of

determination (or squared coefficient of multiple correlation), which is defined as

$$R^2(\mathcal{Z}_n) = 1 - \frac{|\mathbf{C}_n(\{0\} \cup [d])|}{|\mathbf{C}_n([d])|}, \tag{3}$$

where $|\mathbf{C}|$ is the matrix determinant of $\mathbf{C}$ (cf. Ref. 9). Intuitively, the coefficient of determination measures the proportion of the total variation in the response variable that is explained by variation in the covariate vector. See Ref. 7; Sec. 4.4) and Ref. 8; Sec. 3.5) for details regarding the derivation and interpretation of the $R^2(\mathcal{Z}_n)$ coefficients.

2.2. The Shapley Value

A refinement to the question that is addressed by the $R^2(\mathcal{Z}_n)$ coefficient, is that of eliciting the contribution of each of the covariates to the total value of $R^2(\mathcal{Z}_n)$.

In the past, this question has partially been resolved via the use of partial correlation coefficients (see, e.g., Ref. 8; Sec. 3.4). Unfortunately, such coefficients are only able to measure the contribution of each covariate to the coefficient of determination, conditional to the presence of other covariates that are already in the MLR model.

A satisfactory resolution to the question above, is provided by Refs. 10 and 11, and Ref. 1, who each suggested and argued for the use of the Shapley decomposition of Ref. 12.

The Shapley decomposition is a game-theoretic method for decomposing the contribution to the value of a utility function in the context of cooperative games.

Let $\pi^\top = (\pi_1, \ldots, \pi_d)$ be a permutation of the set $[d]$. For each $j \in [d]$, let

$$\mathcal{S}_j(\pi) = \{k : \pi_k < \pi_j, k \in [d]\}$$

be the elements of $[d]$ that appear before π_j when $[d]$ is permuted by π. We may define $R_{\mathcal{S}_j(\pi)}^2(\mathcal{Z}_n)$ and $R_{\{j\} \cup \mathcal{S}_j(\pi)}^2(\mathcal{Z}_n)$ in a similar manner to **Eq. 3**, using the generic definition

$$R_{\mathcal{S}}^2(\mathcal{Z}_n) = 1 - \frac{|\mathbf{C}_n(\{0\} \cup \mathcal{S})|}{|\mathbf{C}_n(\mathcal{S})|}, \tag{4}$$

for nonempty subsets $\mathcal{S} \subseteq [d]$, and $R_{\{\}}^2(\mathcal{Z}_n) = 0$ for the empty set.

Treating the coefficient of determination as a utility function, we may conduct a Shapley partition of the $R^2(\mathcal{Z}_n)$ coefficient by computing the jth Shapley value, for each of the d covariates, defined by

$$V_j(\mathcal{Z}_n) = |\mathcal{P}|^{-1} \sum_{\pi \in \mathcal{P}} \left[R_{\{j\} \cup \mathcal{S}_j(\pi)}^2(\mathcal{Z}_n) - R_{\mathcal{S}_j(\pi)}^2(\mathcal{Z}_n) \right], \tag{5}$$

where $\mathcal{P}$ is the set of all possible permutations of $[d]$.

2.3 Uniqueness of the Shapley Value

Compared to other decompositions of the coefficient of determination, such as those considered in Refs. 13 and 14, the Shapley values, obtained from the partitioning above, have the favorable axiomatic properties that were well exposed in Ref. 1. Specifically, the Shapley values have the efficiency, monotonicity, and, equal treatment properties, and the

decomposition is provably the only method that satisfies all three of these properties (cf. Ref. 15; Thm. 2)). Here, in the context of the coefficient of determination, efficiency, monotonicity, and equal treatment are defined as follows:

Efficiency: The sum of the Shapley values across all covariates equates to the coefficient of determination, that is

$$\sum_{j=1}^{d} V_j(\mathcal{Z}_n) = R^2(\mathcal{Z}_n).$$

Monotonicity: For pairs of samples $\mathcal{Z}_m$ and $\mathcal{Z}_n$, of sizes $m, n \in \mathbb{N}$,

$$R^2_{\{j\} \cup \mathcal{S}_j(\pi)}(\mathcal{Z}_n) - R^2_{\mathcal{S}_j(\pi)}(\mathcal{Z}_n) \geq R^2_{\{j\} \cup \mathcal{S}_j(\pi)}(\mathcal{Z}_m) - R^2_{\mathcal{S}_j(\pi)}(\mathcal{Z}_m),$$

for every $\pi \in \mathcal{P}$, implies that $V_j(\mathcal{Z}_n) \geq V_j(\mathcal{Z}_m)$, for each $j \in [d]$.

Equal treatment: If covariates $j, k \in [d]$ are substitutes in the sense that

$$R^2_{\{j\} \cup \mathcal{S}_j(\pi)}(\mathcal{Z}_n) = R^2_{\{k\} \cup \mathcal{S}_k(\pi)}(\mathcal{Z}_n),$$

for each $\pi \in \mathcal{P}$ such that $k \notin \mathcal{S}_j(\pi)$ and $j \notin \mathcal{S}_k(\pi)$, then the Shapley decomposition is such that $V_j(\mathcal{Z}_n) = V_k(\mathcal{Z}_n)$.

We note that equal treatment is also often referred to as symmetry in the literature. The uniqueness of the Shapley decomposition in exhibiting the three described properties is often used as the justification for its application. Furthermore, there are numerous sets of axiomatic properties that lead to the Shapley value decomposition as a solution (see, e.g., Ref. 16). In the statistics literature, it is known that the axioms for decomposition of the coefficient of determination that are proposed by Ref. 17 correspond exactly to the Shapley values (cf. Ref. 18).

3. THEORETICAL RESULTS

3.1. The Correlation Matrix

Let $Z \in \mathbb{R}^{d+1}$ be a random variable with mean vector $\mu \in \mathbb{R}^{d+1}$ and covariance matrix $\Sigma \in \mathbb{R}^{(d+1) \times (d+1)}$. Then, we can define the coefficient of multivariate kurtosis [19] by

$$\kappa = \frac{1}{(d+1)(d+3)} \left[(Z - \mu)^{\top} \Sigma^{-1} (Z - \mu) \right]^2.$$

Let $\rho_{jk} = \mathrm{cor}(Z_j, Z_k)$, for $j, k \in \{0\} \cup [d]$ such that $j \neq k$. Assume that Z arises from an elliptical distribution (cf. Ref. 20; Ch. 2)) and let $\mathcal{Z}_n$ be an IID sample with the same distribution as Z. Then, we may estimate ρ_{jk} using the sample correlation coefficient **Eq. 1**. Upon writing acov to denote the asymptotic covariance, we have the following result due to Corollary 1 of Ref. 4.

Lemma 1. If Z arises from an elliptical distribution and has coefficient of multivariate kurtosis κ, then the normalized coefficients of correlation $\zeta_{jk} = n^{1/2}(R_{jk} - \rho_{jk})$ $(j, k \in \{0\} \cup [d]; j \neq k)$ converge to a jointly normal distribution with asymptotic mean and covariance elements 0 and

$$\mathrm{acov}(\zeta_{gh}, \zeta_{jk}) = \kappa \left[\rho_{gh} \rho_{jk} \left(\rho_{gj}^2 + \rho_{hj}^2 + \rho_{gk}^2 + \rho_{hk}^2 \right) \middle/ 2 + \rho_{gj} \rho_{hk} + \rho_{gk} \rho_{hj} \right]$$
$$- \kappa \left[\rho_{gh} \left(\rho_{hj} \rho_{hk} + \rho_{gj} \rho_{gk} \right) + \rho_{jk} \left(\rho_{gj} \rho_{hj} + \rho_{gk} \rho_{hk} \right) \right]. \tag{6}$$

Remark 1. We note that the elliptical distribution assumption above can be replaced by a broader pseudo-elliptical assumption, as per Refs. 21 and 22. This is a wide class of distributions that includes some that may not be symmetric. Due to the complicated construction of the class, we refer the interested reader to the source material for its definition.

Remark 2. We may state a similar result that replaces the elliptical assumption by a fourth moments existence assumption instead, using Proposition 2 of Ref. 4. In order to make practical use of such an assumption, we require the estimation of $(d+1)! / [(d-3)!4!]$ fourth order moments instead of a single kurtosis term κ. Such a result may be useful when the number of fourth order moments is small, but become infeasible rapidly, as d increases.

Let $v \subseteq \{0\} \cup [d]$, where $v = \{v_1, \ldots, v_{|v|}\}$. Define $C_n(v)$ in the same manner as **Eq. 2**, and let

$$\mathbf{R}(\mathcal{U}) = \begin{bmatrix} 1 & \rho_{u_1 u_2} & \cdots & \rho_{u_1 u_{|\mathcal{U}|}} \\ \rho_{u_2 u_1} & 1 & \cdots & \rho_{u_2 u_{|\mathcal{U}|}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{u_{|\mathcal{U}|} u_1} & \rho_{u_{|\mathcal{U}|} u_2} & \cdots & 1 \end{bmatrix}$$

and

$$\mathbf{R}(v) = \begin{bmatrix} 1 & \rho_{v_1 v_2} & \cdots & \rho_{v_1 v_{|v|}} \\ \rho_{v_2 v_1} & 1 & \cdots & \rho_{v_2 v_{|v|}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{v_{|v|} v_1} & \rho_{v_{|v|} v_2} & \cdots & 1 \end{bmatrix}.$$

The following theorem is adapted from a result of Ref. 5 (also appearing as Theorem 1 in Ref. 23). Our result expands upon the original theorem, to allow for inference regarding elliptically distributed data, and not just normally distributed data. We further fix some typographical matters that appear in both Refs. 5 and 23.

Lemma 2. *Assume the same conditions as in* **Eq. 1.** *Then, the normalized covariance determinant* $\delta(\mathcal{U}) = n^{1/2}(|C_n(\mathcal{U})| - |\mathbf{R}(\mathcal{U})|)$ *(where $\mathcal{U}$ and v are nonempty subsets of $\{0\} \cup [d]$) converges to a jointly normal distribution, with asymptotic mean and covariance elements 0 and*

$$\mathrm{acov}(\delta(\mathcal{U}), \delta(v)) = \sum_{g,h \in \mathcal{U}} \sum_{j,k \in v} r_{\mathcal{U}}(g,h) r_v(j,k) \mathrm{acov}(\zeta_{gh}, \zeta_{jk}), \tag{7}$$

where $\mathrm{acov}(\zeta_{gh}, \zeta_{jk})$ *is as per* **Eq. 6,**

$$\begin{bmatrix} r_{\mathcal{U}}(u_1, u_1) & r_{\mathcal{U}}(u_1, u_2) & \cdots & r_{\mathcal{U}}(u_1, u_{|\mathcal{U}|}) \\ r_{\mathcal{U}}(u_2, u_1) & r_{\mathcal{U}}(u_2, u_2) & \cdots & r_{\mathcal{U}}(u_2, u_{|\mathcal{U}|}) \\ \vdots & \vdots & \ddots & \vdots \\ r_{\mathcal{U}}(u_{|\mathcal{U}|}, u_1) & r_{\mathcal{U}}(u_{|\mathcal{U}|}, u_2) & \cdots & r_{\mathcal{U}}(u_{|\mathcal{U}|}, u_{|\mathcal{U}|}) \end{bmatrix}$$
$$= |\mathbf{R}(\mathcal{U})| \mathbf{R}^{-1}(\mathcal{U}),$$

and $r_v(j,k)$ $(j, k \in v)$ *is defined similarly.*

Proof. The result is due to an application of the delta method (see, e.g., Ref. 24; Thm. 3.1)) and the fact that for any matrix $\mathbf{R}$, the derivative of its determinant is $\partial|\mathbf{R}|/\partial\mathbf{R} = |\mathbf{R}|\mathbf{R}^{-\top}$ [25]; Sec. 17.45). Notice that we use the unconstrained case of the determinant derivative, since we sum over each pair of coordinates, where $g \neq h$ or $j \neq k$, twice. ∎

Remark 3. If $\mathbf{R}$ is symmetric, then $\partial|\mathbf{R}|/\partial\mathbf{R} = |\mathbf{R}|[2\mathbf{R}^{-1} - diag(\mathbf{R}^{-1})]$ [25]; Sec. 17.45). Using this fact, we may write **Eq. 7** in the alternative, and more computationally efficient form

$$acov(\delta(\mathcal{U}), \delta(v)) = \sum_{\substack{g \leq h \\ g,h \in \mathcal{U}}} \sum_{\substack{j \leq k \\ j,k \in v}} r_\mathcal{U}^\star(g,h) r_v^\star(j,k) acov(\zeta_{gh}, \zeta_{jk}),$$

where

$$\begin{bmatrix} r_\mathcal{U}^\star(u_1,u_1) & r_\mathcal{U}^\star(u_1,u_2) & \cdots & r_\mathcal{U}^\star(u_1,u_{|\mathcal{U}|}) \\ r_\mathcal{U}^\star(u_2,u_1) & r_\mathcal{U}^\star(u_2,u_2) & \cdots & r_\mathcal{U}^\star(u_2,u_{|\mathcal{U}|}) \\ \vdots & \vdots & \ddots & \vdots \\ r_\mathcal{U}^\star(u_{|\mathcal{U}|},u_1) & r_\mathcal{U}^\star(u_{|\mathcal{U}|},u_2) & \cdots & r_\mathcal{U}^\star(u_{|\mathcal{U}|},u_{|\mathcal{U}|}) \end{bmatrix}$$
$$= 2|\mathbf{R}(\mathcal{U})|\mathbf{R}^{-1}(\mathcal{U})$$
$$- |\mathbf{R}(\mathcal{U})|diag(\mathbf{R}^{-1}(\mathcal{U})),$$

and $r_v^\star(j,k)$ $(j,k \in v)$ is defined similarly.

3.2. The Coefficient of Determination

Let plim denote convergence in probability, so that for any sequence $\{X_n\}$, and any random variable X, the statement

$$\lim_{n \to \infty} \Pr(|X_n - X| > \varepsilon) = 0,$$

for every $\varepsilon > 0$, can be written as $\text{plim}_{n \to \infty} X_n = X$.

Now, recall definition **Eq. 4**, and further let $\rho_\mathcal{S}^2 = \text{plim}_{n \to \infty} R_\mathcal{S}^2(\mathcal{Z}_n)$. We adapt from and expand upon [23]; Thm. 2) in the following result. This result also fixes typographical errors that appear in the original theorem, as well as in Ref. 5.

Lemma 3. *Assume the same conditions as in* **Eq. 1**. *Then, the normalize coefficient of determination* $\lambda(\mathcal{S}) = n^{1/2}(R_\mathcal{S}^2(\mathcal{Z}_n) - \rho_\mathcal{S}^2)$ *(where S and* $\mathcal{T}$ *are nonempty subsets of* $[d]$*) converges to a jointly normal distribution, with asymptotic mean and covariance elements 0 and*

$$acov(\lambda(\mathcal{S}), \lambda(\mathcal{T})) = \frac{1}{|\mathbf{R}(\mathcal{S})||\mathbf{R}(\mathcal{T})|} acov(\delta(\{0\}\cup\mathcal{S}), \delta(\{0\}\cup\mathcal{T}))$$
$$+ \frac{|\mathbf{R}(\{0\}\cup\mathcal{S})||\mathbf{R}(\{0\}\cup\mathcal{T})|}{|\mathbf{R}(\mathcal{S})|^2|\mathbf{R}(\mathcal{T})|^2} acov(\delta(\mathcal{S}), \delta(\mathcal{T}))$$
$$- \frac{|\mathbf{R}(\{0\}\cup\mathcal{S})|}{|\mathbf{R}(\mathcal{S})|^2|\mathbf{R}(\mathcal{T})|} acov(\delta(\mathcal{S}), \delta(\{0\}\cup\mathcal{T}))$$
$$- \frac{|\mathbf{R}(\{0\}\cup\mathcal{T})|}{|\mathbf{R}(\mathcal{S})||\mathbf{R}(\mathcal{T})|^2} acov(\delta(\{0\}\cup\mathcal{S}), \delta(\mathcal{T})). \tag{8}$$

Proof. We apply the delta method again, using the functional form **Eq. 4**, and using the fact that

$$\frac{\partial}{\partial(x,y)}\left(1 - \frac{x}{y}\right) = \left(-\frac{1}{y}, \frac{x}{y^2}\right).$$

Remark 4. When $\mathcal{S} = \mathcal{T}$, **Eq. 8** yields the usual form for the asymptotic variance of $R_\mathcal{S}^2(\mathcal{Z}_n)$: $4\kappa\rho_\mathcal{S}^2(1 - \rho_\mathcal{S}^2)^2$ (cf. Ref. 22).

3.3. A Shapley Value Confidence Interval

For every $j \in [d]$ and $\mathcal{S} \subseteq \mathcal{S}_j = [d] - \{j\}$, there are $|\mathcal{S}|!(d - |\mathcal{S}| - 1)!$ elements of $\pi \in \mathcal{P}$ such that $\mathcal{S}_j(\pi) = \mathcal{S}$. Thus, we may write

$$V_j(\mathcal{Z}_n) = \sum_{\mathcal{S} \subseteq \mathcal{S}_j} \omega(\mathcal{S})\left[R_{\{j\}\cup\mathcal{S}}^2(\mathcal{Z}_n) - R_\mathcal{S}^2(\mathcal{Z}_n)\right], \tag{9}$$

where $\omega(\mathcal{S}) = |\mathcal{S}|!(d - |\mathcal{S}| - 1)!/d!$, and define $v_j = \text{plim}_{n \to \infty} V_j(\mathcal{Z}_n)$. Using this functional form **Eq. 9**, we may apply the delta method once more, in order to derive the following joint asymptotic normal distribution result regarding the Shapley values $V_j(\mathcal{Z}_n)$, for $j \in [d]$.

Remark 5. The form **Eq. 9** is a useful computational trick that reduces the computational time of form **Eq. 5** and results in more efficient computations for fixed d. It is unclear whether other formulations such as that of Ref. 26 can make the computation time even faster. Unfortunately, however, there is no formulation that reduces the $\mathcal{O}(2^d)$ scaling, as d increases.

Theorem 1. Assume the same conditions as in **Eq. 1**. Then, the normalized Shapley values $\xi_j = n^{1/2}(V_j(\mathcal{Z}_n) - v_j)$ (where $j,k \in [d]$) converge to a jointly normal distribution, with asymptotic mean and covariance elements 0 and $acov(\xi_j, \xi_k) = a_{jk} + b_{jk} - c_{jk} - d_{jk}$. Here,

$$a_{jk} = \sum_{\mathcal{S} \subseteq \mathcal{S}_j} \sum_{\mathcal{T} \subseteq \mathcal{S}_k} \omega(\mathcal{S})\omega(\mathcal{T}) acov(\lambda(\{j\}\cup\mathcal{S}), \lambda(\{k\}\cup\mathcal{T})),$$

$$b_{jk} = \sum_{\mathcal{S} \subseteq \mathcal{S}_j} \sum_{\mathcal{T} \subseteq \mathcal{S}_k} \omega(\mathcal{S})\omega(\mathcal{T}) acov(\lambda(\mathcal{S}), \lambda(\mathcal{T})),$$

$$c_{jk} = \sum_{\mathcal{S} \subseteq \mathcal{S}_j} \sum_{\mathcal{T} \subseteq \mathcal{S}_k} \omega(\mathcal{S})\omega(\mathcal{T}) acov(\lambda(\mathcal{S}), \lambda(\{k\}\cup\mathcal{T})),$$

and

$$d_{jk} = \sum_{\mathcal{S} \subseteq \mathcal{S}_j} \sum_{\mathcal{T} \subseteq \mathcal{S}_k} \omega(\mathcal{S})\omega(\mathcal{T}) acov(\lambda(\{j\}\cup\mathcal{S}), \lambda(\mathcal{T})),$$

where $\lambda(\mathcal{S})$ is as defined in **Eq. 3**, for nonempty subsets $\mathcal{S} \subseteq [d]$.

Using the result above, we may apply the delta method again in order to construct asymptotic CIs or hypothesis tests regarding any continuous function of the d Shapley values for the coefficient of determination. Of particular interest is the asymptotic CI for each of the individual Shapley values and the hypothesis test for the difference between two Shapley values.

The asymptotic $100(1 - \alpha)\%$ CI for the jth expected Shapley value v_j has the usual form

$$\left(V_j(\mathcal{Z}_n) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\text{avar}(\xi_j)}{n}}, V_j(\mathcal{Z}_n)\right.$$
$$\left. + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\text{avar}(\xi_j)}{n}}\right),$$

where $\mathrm{avar}(\xi_j) = \mathrm{acov}(\xi_j, \xi_j)$ denotes the asymptotic variance of ξ_j and Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution. The z-statistic for the test of the null and alternative hypotheses

$$H_0 : \nu_j = \nu_k \text{ and } H_1 : \nu_j \neq \nu_k, \tag{10}$$

for $j, k \in [d]$ such that $j \neq k$, is

$$\Delta_n = \frac{\sqrt{n}\big(V_j(\mathscr{Z}_n) - V_k(\mathscr{Z}_n)\big)}{\sqrt{\mathrm{avar}(\xi_j) + \mathrm{avar}(\xi_k) - 2\mathrm{acov}(\xi_j, \xi_k)}},$$

where Δ_n has an asymptotic standard normal distribution.

Remark 6. In practice, we do not know the necessary elements κ and ρ_{jk} $(j, k \in \{0\} \cup [d]; j \neq k)$ that are required in order to specify the asymptotic covariance terms in **Eqs 1–3** and **Eq. 1**. However, by Slutsky's theorem, we have the usual result that any acov (or avar) term can be replaced by the estimator $\widehat{acov}_n$ (or $\widehat{avar}_n$), which replaces κ by the estimator of Ref. [19]

$$\hat{\kappa}(\mathscr{Z}_n) = \frac{\sum_{i=1}^{n} \left[(\mathbf{Z}_i - \overline{\mathbf{Z}})^\top \hat{\Sigma}_n^{-1}(\{0\} \cup [d])(\mathbf{Z}_i - \overline{\mathbf{Z}}) \right]^2}{n(d+1)(d+3)},$$

and replaces ρ_{jk} by $R_{jk}(\mathscr{Z}_n)$, where $\overline{\mathbf{Z}}^\top = (\overline{Z}_0, \dots, \overline{Z}_d)$. Here,

$$\hat{\Sigma}_n(\mathcal{U}) = \begin{bmatrix} \hat{\Sigma}_{u_1 u_1}(\mathscr{Z}_n) & \hat{\Sigma}_{u_1 u_2}(\mathscr{Z}_n) & \cdots & \hat{\Sigma}_{u_1 u_{|\mathcal{U}|}}(\mathscr{Z}_n) \\ \hat{\Sigma}_{u_2 u_1}(\mathscr{Z}_n) & \hat{\Sigma}_{u_2 u_2}(\mathscr{Z}_n) & \cdots & \hat{\Sigma}_{u_2 u_{|\mathcal{U}|}}(\mathscr{Z}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Sigma}_{u_{|\mathcal{U}|} u_1}(\mathscr{Z}_n) & \hat{\Sigma}_{u_{|\mathcal{U}|} u_2}(\mathscr{Z}_n) & \cdots & \hat{\Sigma}_{u_{|\mathcal{U}|} u_{|\mathcal{U}|}}(\mathscr{Z}_n) \end{bmatrix},$$

where $\hat{\Sigma}_{jk}(\mathscr{Z}_n)$ is the sample covariance between the jth and kth dimension of $\mathbf{Z}$ $(j, k \in \{0\} \cup [d])$. For example, the estimated test statistic

$$\hat{\Delta}_n = \frac{\sqrt{n}\big(V_j(\mathscr{Z}_n) - V_k(\mathscr{Z}_n)\big)}{\sqrt{\widehat{avar}_n(\xi_j) + \widehat{avar}_n(\xi_k) - 2\widehat{acov}_n(\xi_j, \xi_k)}}, \tag{11}$$

for the hypotheses **Eq. 10**, retains the property of having an asymptotically standard normal distribution.

The confidence interval calculation is summarized in Algorithm 1.

Algorithm 1: A Shapley value confidence interval
1. **Input**: dataset $\mathscr{Z}_n$, significance level α, and feature index j
2. Compute the estimate $V_j(\mathscr{Z}_n)$ of the Shapley value ν_j, via **Eq. 9**
3. Compute $a_{jj}, b_{jj}, c_{jj}, d_{jj}$ from Theorem 1, taking care to replace ρ_{jk} in Lemmas 1–3 by the estimate $R_{jk}(\mathscr{Z}_n)$ (see Remark 6)
4. Compute $\widehat{avar}_n(\xi_j) = a_{jj} + b_{jj} - c_{jj} - d_{jj}$
5. **Output**: a $100(1-\alpha)\%$ confidence interval for ν_j is then given by

$$\left(V_j(\mathscr{Z}_n) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\widehat{avar}_n(\xi_j)}{n}}, V_j(\mathscr{Z}_n) \right.$$
$$\left. + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\widehat{avar}_n(\xi_j)}{n}} \right)$$

4. MONTE CARLO STUDIES AND BENCHMARKS

In each of the following three Monte Carlo studies, we simulate a large number N of random samples $\mathscr{Z}_n^{(i)}$, $i \in [N]$, of size n, from a chosen distribution $\mathcal{D}$. For each sample, we apply **Eq. 1** to calculate an asymptotic 95% CI for the first Shapley value ν_1, producing a set of N observed intervals $\mathcal{I}_N = \{[\ell_i, u_i] : i \in [N]\}$, as realisations of the CI $[L, U]$ for ν_1. The coverage probability $\mathrm{covr}_{\nu_1}([L, U]) := \mathrm{Pr}(\nu_1 \in [L, U])$ is then estimated as the proportion of intervals in $\mathcal{I}_N$ containing the population Shapley value

$$\widehat{\mathrm{covr}}_{\nu_1}([L, U]) = \frac{|\{[\ell_i, u_i] \in \mathcal{I}_N : \nu_1 \in [\ell_i, u_i]\}|}{N}. \tag{12}$$

Here, the population Shapley value ν_1 has the form:

$$\nu_1 = \sum_{\mathcal{S} \subseteq \mathcal{S}_1} \omega(\mathcal{S})\big[R^2_{\{1\} \cup \mathcal{S}} - R^2_\mathcal{S}\big],$$

where $R^2_\mathcal{S}$ is defined by replacing $\mathbf{C}_n$ in **Eq. 4** by the known population correlation matrix $\mathrm{cor}(\mathbf{Z})$, as determined by the chosen distribution $\mathcal{D}$. In Studies A and B, this population correlation matrix is the $(d+1) \times (d+1)$ matrix Σ, with diagonal elements equal to 1 and off-diagonal elements equal to a constant correlation $c \in [0, 1)$. That is,

$$\Sigma = c\mathbf{J}_{d+1} + (1-c)\mathbf{I}_{d+1}, \tag{13}$$

where $\mathbf{J}_{d+1}$ denotes a $(d+1) \times (d+1)$ matrix with all entries equal to 1. Note that, for fixed variance, larger magnitudes of c map to larger regression coefficients. Thus, these simulations can be viewed as assigning equal regression coefficients to each covariates that are increasing for increasing values of c.

In Study C, we are concerned with covariance matrices with off-diagonal elements deviating from c. We aim to capture a case where the off-diagonal elements of $\mathrm{cor}(\mathbf{Z})$ are not uniform, and where some may be negative. This is achieved by sampling $\mathscr{Z}_n^{(i)}$ from a multivariate normal distribution $\mathcal{D}_i$, with a symmetric positive definite covariance matrix Σ_i that is sampled at random from a Wishart distribution with scale matrix Σ; see **Section 4.3**. Accordingly, the population Shapley value $\nu_1^{(i)}$ is unique to sample i, and thus we adjust the coverage estimator $\widehat{\mathrm{covr}}_{\nu_1}([L, U])$ by replacing ν_1 on the RHS of **Eq. 12** by $\nu_1^{(i)}$.

To accompany our estimates of $\widehat{\mathrm{covr}}_{\nu_1}([L, U])$, we also provide Clopper-Pearson CIs [27] for the coverage probability. We also report the average CI widths, and middle 95% percentile intervals for the widths. For comparison, we estimate coverage probabilities of non-parametric bootstrap confidence intervals in each of the three studies. To obtain the bootstrap CIs, we set some large number N_b and take random resamples $\mathcal{R}_n^{(r)}$, $r \in [N_b]$, of size n, with replacement, from $\mathscr{Z}_n^{(i)}$. From these resamples, we calculate the set of estimated Shapley values $\mathcal{L}_i = \{V_1(\mathcal{R}_n^{(r)}) : r \in [N_b]\}$. The ith 95% bootstrap CI is then taken as the middle 95% percentile interval of $\mathcal{L}$, and the coverage is estimated as in **Eq. 12**.
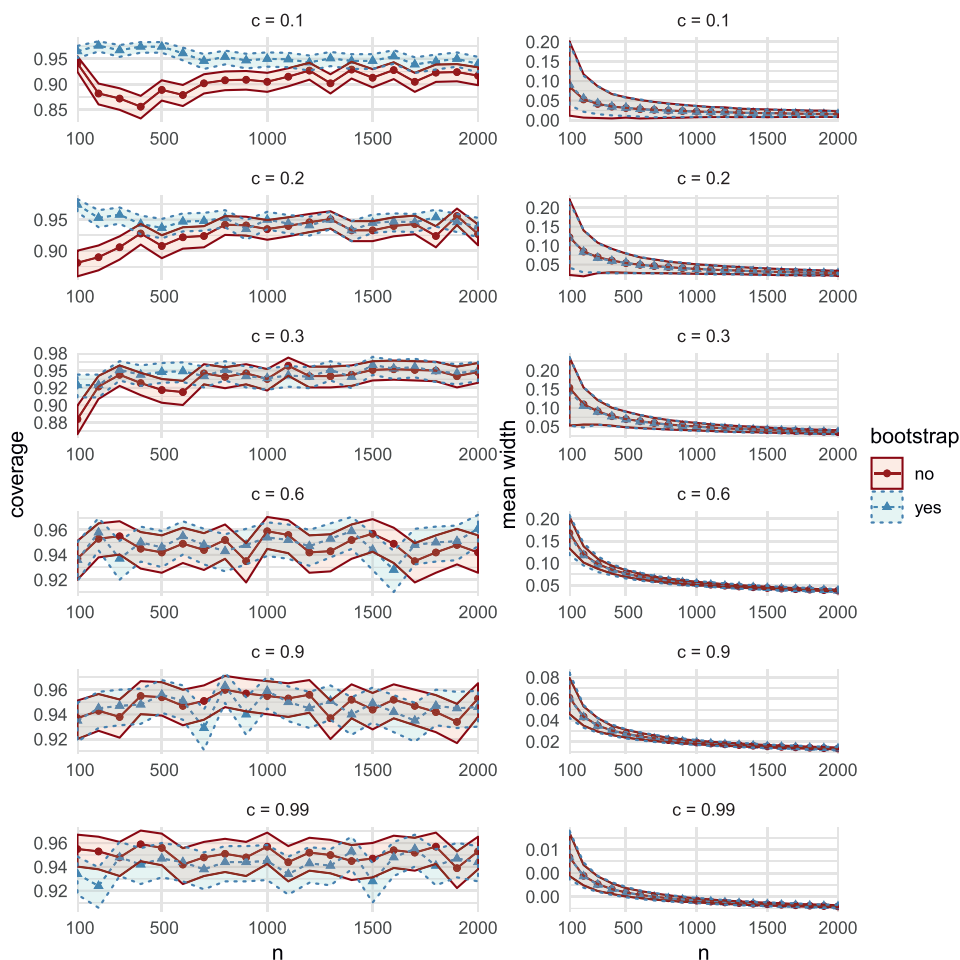
FIGURE 1 | Comparisons of coverage **(left column)** and mean width **(right column)**, between the bootstrap CIs (dashed lines with triangular markers) and the asymptotic CIs (solid lines with circular markers) in MC Study A. Rows represent correlations c, increasing from 0.1 on the top row to 0.99 on the bottom row. The horizontal axes display the sample sizes $n = 100, 200, \dots, 2000$.

To obtain results for each pair $(n, c) \in \mathcal{N} \times \mathcal{C}$, where $\mathcal{N} = \{5, 10, \dots, 50\} \cup \{100, 200, \dots, 2000\}$ and $\mathcal{C} = \{0, 0.1, 0.2, 0.3, 0.6, 0.9, 0.99\}$, we performed $30 \times 7 = 210$ simulations for each of the three studies. We use $N = N_b = 1000$ and $d = 3$, in all cases.

4.1. MC Study A

Here, we choose $\mathcal{D} = N_{d+1}(0, \Sigma)$, so that each sample $\mathcal{Z}_n^{(i)}$, for $i \in [N]$, is drawn from a multivariate normal distribution, with covariance Σ given in **Eq. 13**.

The simulation results in **Figure 1** (and in **Supplementary Figure S12**) show very similar coverage and width performance between the two assessed CIs for moderate and high correlations $c > 0.3$. For lower correlations $c \leq 0.2$, coverage convergence appears to be slower in n than the bootstrap CI for large sample sizes ($n \geq 100$). The opposite trend seems to hold for small sample sizes ($n \leq 50$), see the discussion under MC Study C. Also, for the highest correlation $c = 0.99$, coverage performance of the asymptotic CI is overall slightly better than the bootstrap CI.

4.2. MC Study B

Here, we choose $\mathcal{D} = t_\nu(0, \Sigma)$ where $t_\nu(\mu, \Sigma)$ is the multivariate Student t distribution with $\nu \in (0, \infty)$ degrees of freedom, mean vector μ and scale matrix Σ. Specifically, the ith sample $\mathcal{Z}_n^{(i)}$, for $i \in [N]$, we set $\nu = 100$ degrees of freedom, and set Σ as the $(d + 1) \times (d + 1)$ covariance matrix in **Eq. 13**.

For all sample sizes n and correlations c, coverage and width performances are similar to MC Study A (see **Supplementary Figures S13 and S14**). Of particular interest, in both MC Studies A and B (but not in MC Study C), we observe that for $c = 0$, the estimated coverage probability of the asymptotic CI is almost equal to 1, for all sample sizes greater than 10, while the corresponding bootstrap CIs have estimated coverage equal to 0 (**Figure 2** left). Despite this, the average CI widths, though large under small samples, are somewhat smaller than those for bootstrap (**Figure 2** right).

4.3. MC Study C

Here, we set $\mathcal{D}_i = N_{d+1}(0, \Sigma_i)$, so that the sample $\mathcal{Z}_n^{(i)}$, for $i \in N$, is drawn from a multivariate normal distribution, with covariance
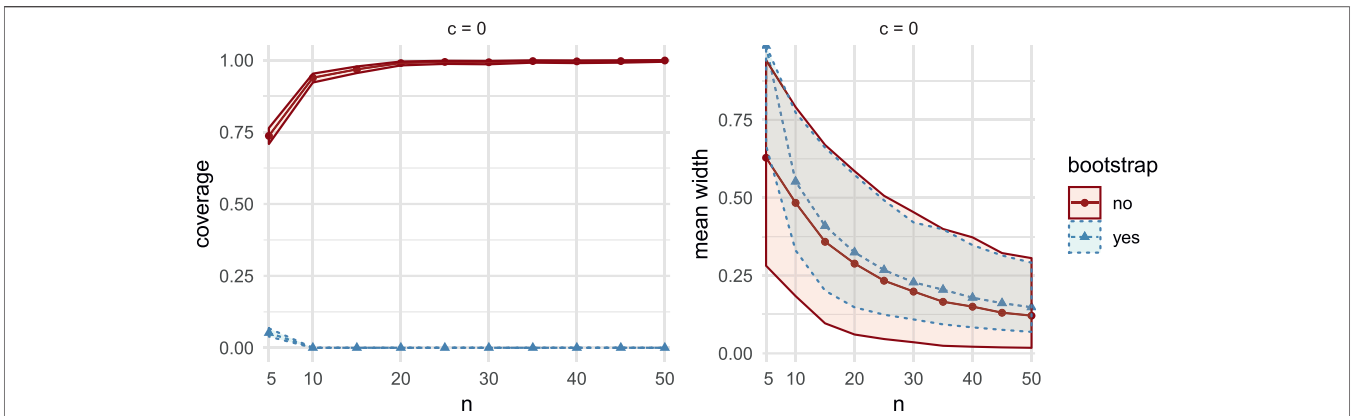
FIGURE 2 | Comparisons of coverage **(left column)** and mean width **(right column)** between bootstrap CIs (dashed lines with triangular markers) and the asymptotic CIs (solid lines with circular markers) in MC Study B, for correlation $c = 0$. The horizontal axes display the sample sizes $n = 5, 10, \ldots, 50$.
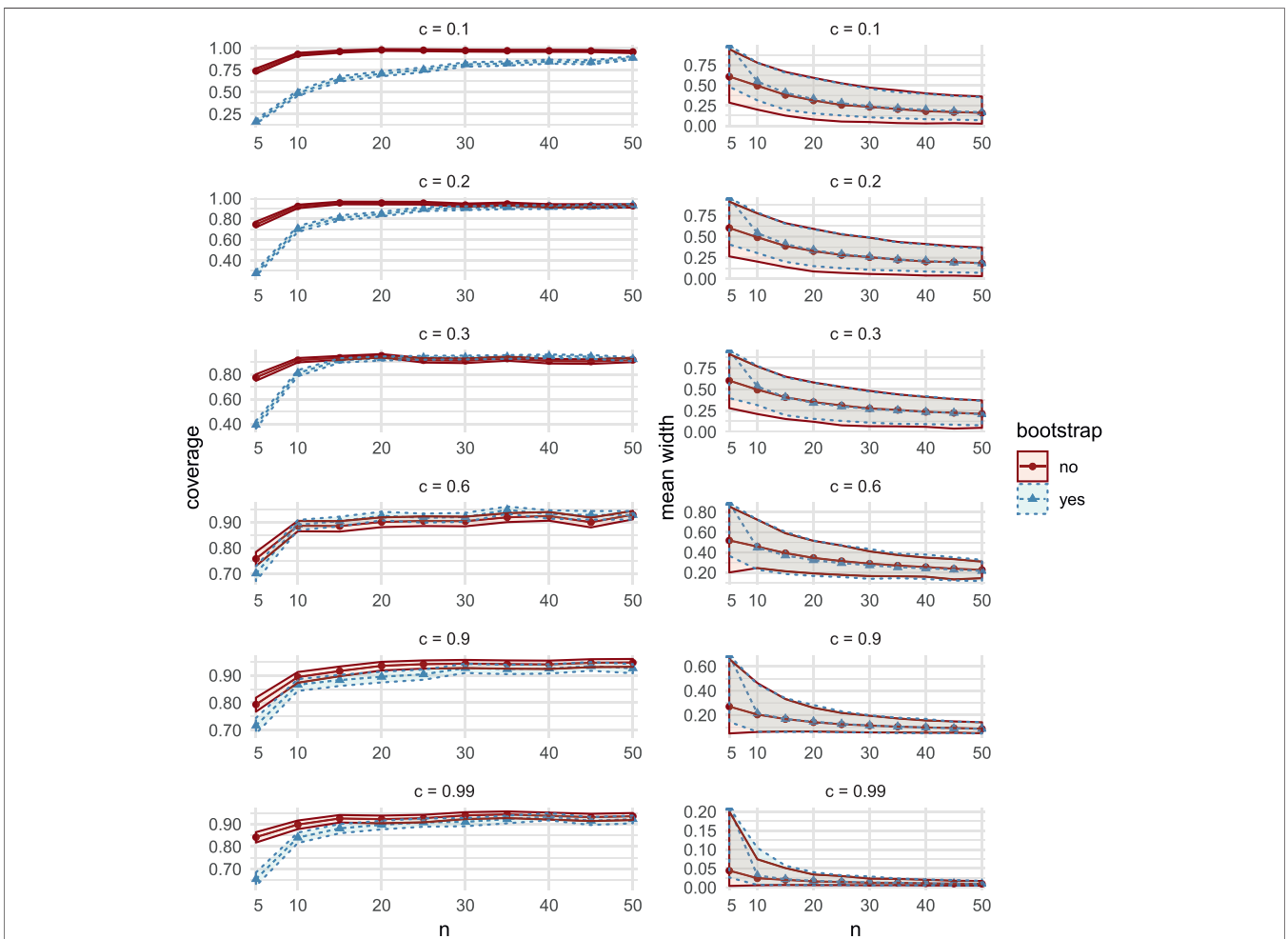

FIGURE 3 | Comparisons of coverage **(left column)** and mean width **(right column)**, between the bootstrap CIs (dashed lines with triangular markers) and the asymptotic CIs (solid lines with circular markers) in MC Study C. Rows represent correlations c, increasing from 0.1 on the top row to 0.99 on the bottom row. The horizontal axes display the sample sizes $n = 5, 10, \ldots, 50$.
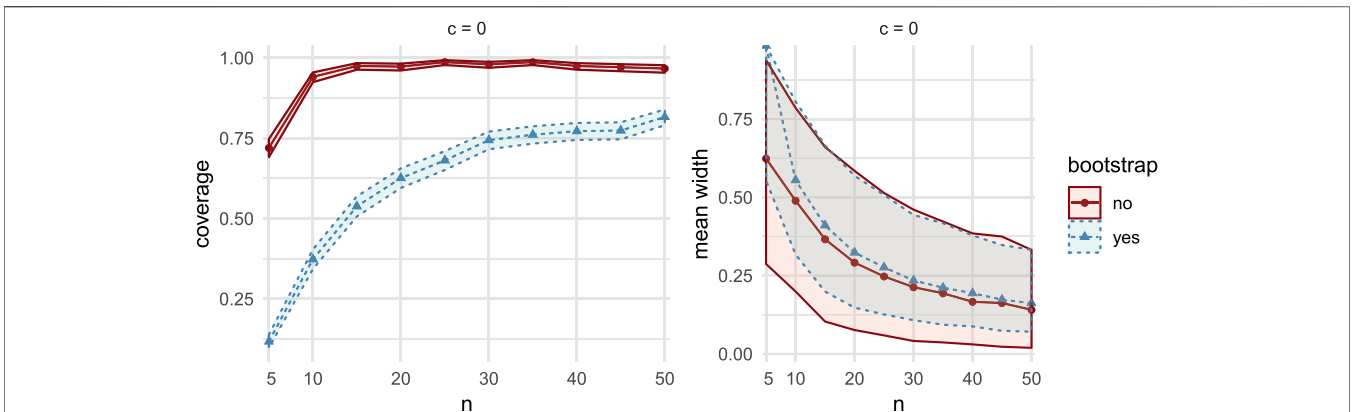
FIGURE 4 | Comparisons of coverage **(left column)** and mean width **(right column)** between bootstrap CIs (dashed lines with triangular markers) and the asymptotic CIs (solid lines with circular markers) in MC Study C, for correlation $c = 0$. The horizontal axes display the sample sizes $n = 5, 10, \ldots, 50$.
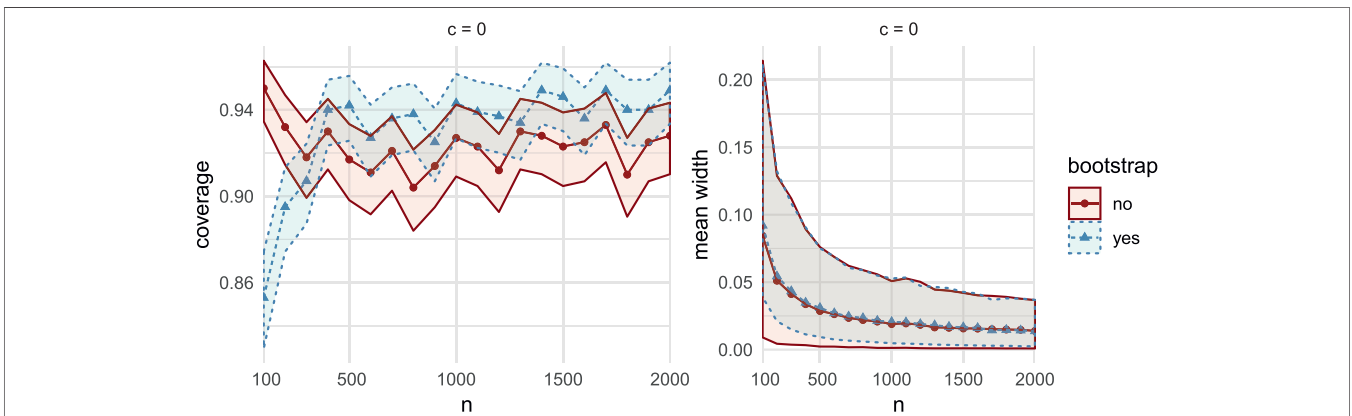

FIGURE 5 | Comparisons of coverage **(left column)** and mean width **(right column)** between bootstrap CIs (dashed lines with triangular markers) and the asymptotic CIs (solid lines with circular markers) in MC Study C, for correlation $c = 0$. The horizontal axes display the sample sizes $n = 100, 200, \ldots, 2000$.
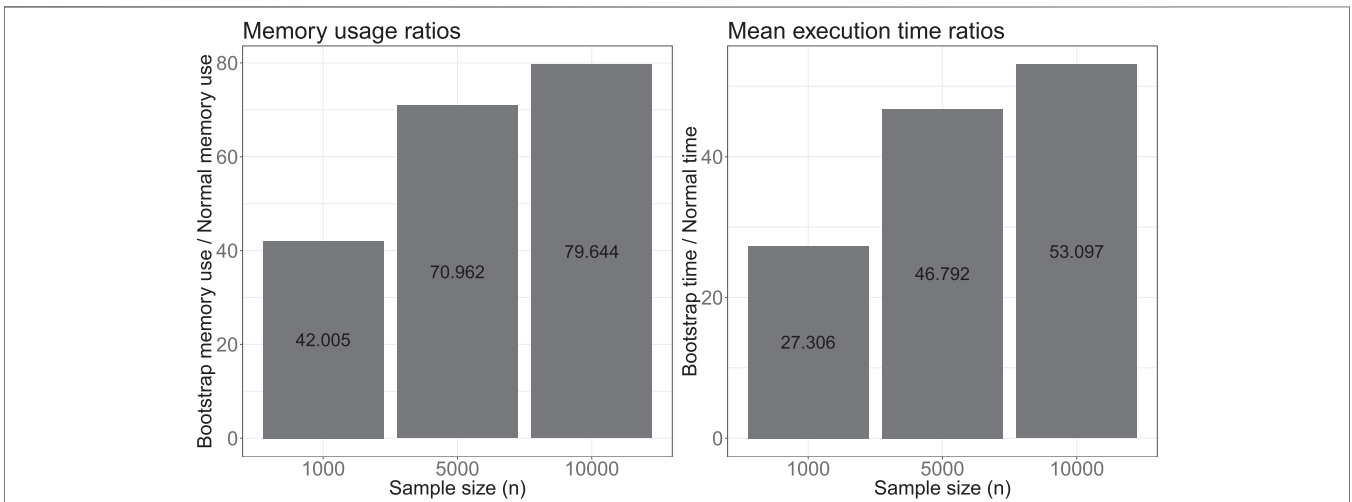

FIGURE 6 | Computational benchmark metric ratios of confidence interval estimation using the naïve bootstrap over the asymptotic normality approach.
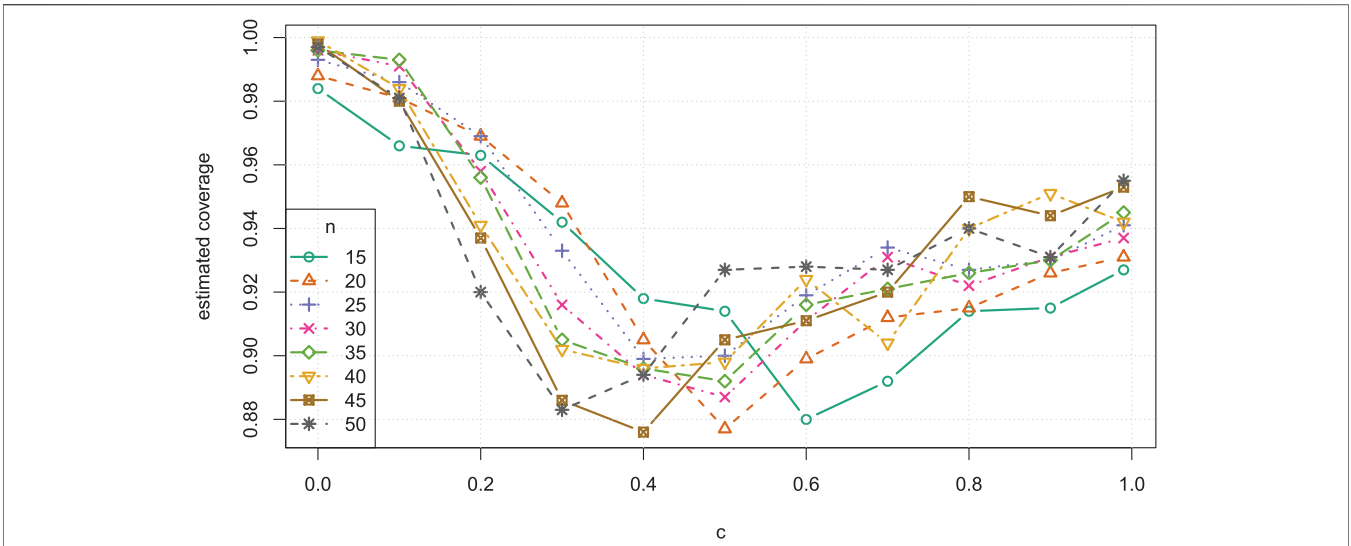
FIGURE 7 | Estimated coverage probability versus correlation c, for small samples sizes $n = 15, 20, \ldots, 50$, in MC Study A. The same patterns can be observed for Studies B and C.
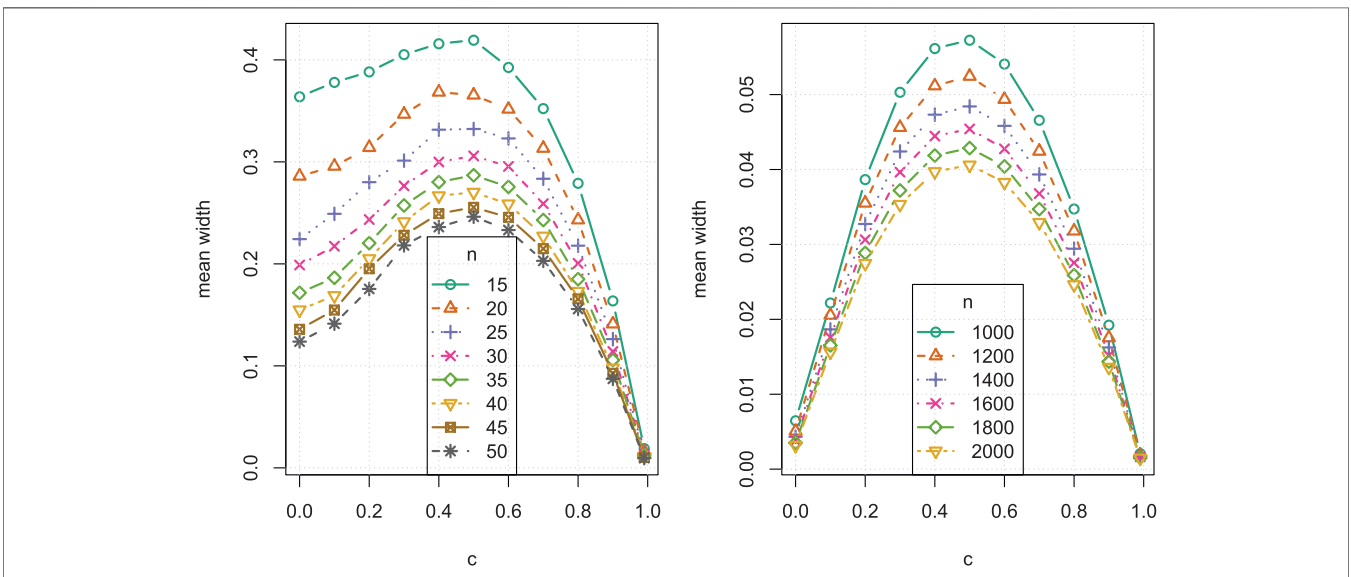

FIGURE 8 | Average confidence interval width versus correlation c, for small sample sizes $n = 15, 20, \ldots, 50$ (left) and large sample sizes $n = 1000, 1200, \ldots, 2000$ (right), in MC Study A. The same patterns are present in MC Study B and C.

matrix Σ_i realized from a Wishart distribution $W_{d+1}(\Sigma, \nu)$ with scale matrix Σ and ν degrees of freedom. This set up is different from Studies A and B in that the distributions $\mathcal{D}_i$, and therefore the population Shapley values $v_1^{(i)}$, are allowed to differ between samples.

The distribution $W_{d+1}(\Sigma, \nu)$ can be understood as the distribution of the sample covariance matrix of a sample of size $\nu + 1$ from the distribution $N_{d+1}(0, \Sigma)$ (cf. Ref. 28). This implies that each covariance matrix Σ_i can have non-uniform and negative off-diagonal elements, with variability between the

off-diagonal elements increasing as ν decreases. For this study, we set $\nu = 100$.

Aside for the case $c = 0$, coverage and width statistics are again similar to MC Studies A and B, for all n and c (see **Supplementary Material**). Interestingly, in all three Studies, for small sample sizes ($n \leq 50$), coverage is often higher than for the bootstrap CI, with slightly smaller average widths, as seen in **Figure 3** (and in **Supplementary Figures S12, S13**). For the $c = 0$ case, the observed behavior differs from MC Studies A and B, with bootstrap performing comparatively well for large sample sizes (**Figures 4, 5**).
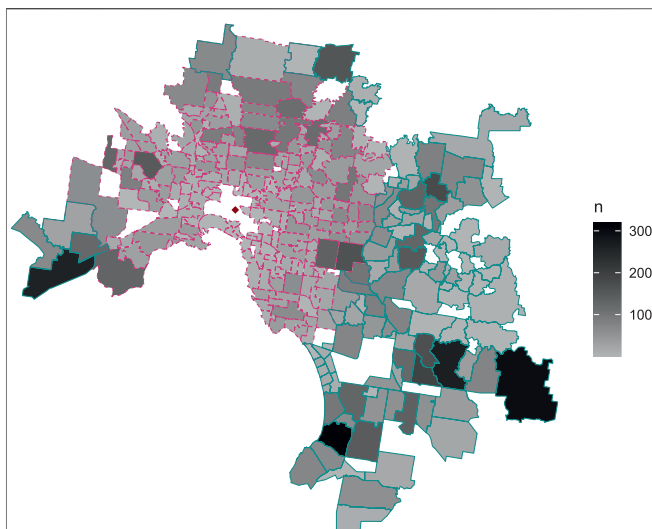
FIGURE 9 | Map of Melbourne suburbs included in this study, with greyscale shading to represent sample sizes. Pink dashed polygon borders contain suburbs classified as near to the Central Business District (i.e., ≤ 25 km), and blue solid polygon borders represent those classified as far (> 25 km). The solid red diamond represents the location of the CBD.
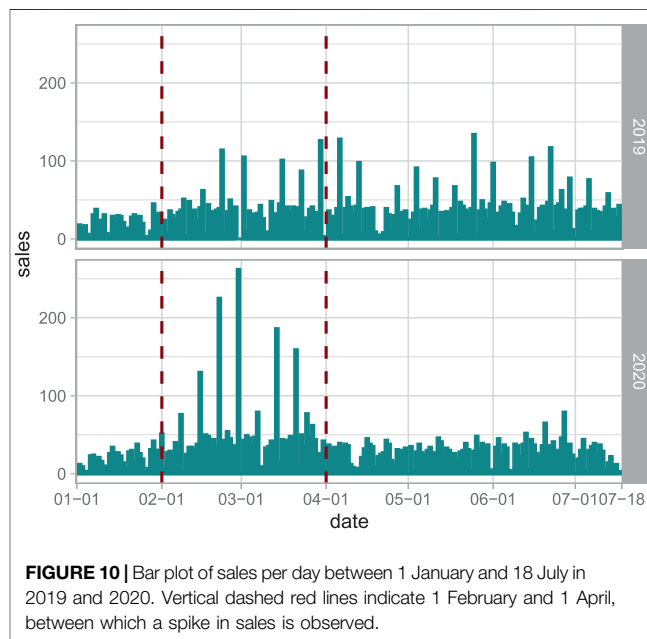

FIGURE 10 | Bar plot of sales per day between 1 January and 18 July in 2019 and 2020. Vertical dashed red lines indicate 1 February and 1 April, between which a spike in sales is observed.

4.4. Computational Benchmarks

From **Figure 6**, we see that the memory usage (left) and mean execution time (right) for the bootstrap CIs are both higher than that for the asymptotic CIs, and that the ratio increases with sample size given the benchmarking parameters in **Table 1**. As n increases, asymptotic CIs become increasingly efficient, compared to the bootstrap CIs. On the other hand, as d increases, with n fixed, we expect an increase in the relative efficiency of the bootstrap, since the complexity of calculating acov(ξ_j, ξ_k) in **Eq. 1** grows faster in d than the complexity of the bootstrap procedure.

4.5. Summary of Results and Recommendations for Use

The most relevant observation is that the widths all appear to shrink at a rate that is comparable to the expected $\sqrt{n}$ convergence dictated by asymptotic normality. The remaining observations below are focused on computational time, sample size, and correlation, in comparison with bootstrap intervals. Below follows a summary of the general tendencies and observations from our results, and recommendations regarding when to use the asymptotic CIs.

- For all correlations c in all three Studies, the estimated coverage probability of asymptotic intervals is above 0.85 for all sample sizes $n \geq 10$.
- For smaller correlations and sample sizes, in particular $c \geq 0.2$ and $n > 15$, the lower bound of the confidence interval for coverage never drops below 0.85.
- For all correlations $c \geq 0.3$ and sample sizes $n > 100$, the lower bound of the confidence interval for coverage never drops below 0.9.

TABLE 1 | Parameters for computational benchmarking.

Parameter	Value(s)
Number of features (d)	3
Sample sizes (n)	1,000, 5,000, 10,000
Number of bootstrap resamples (N_b)	1,000
Number of simulation repetitions (N)	1,000

TABLE 2 | The four subgroups and their sample sizes after partitioning by distance (where near:= within 25 km of CBD), and year of sale in the period 1 February to 1 April 2019 and 2020.
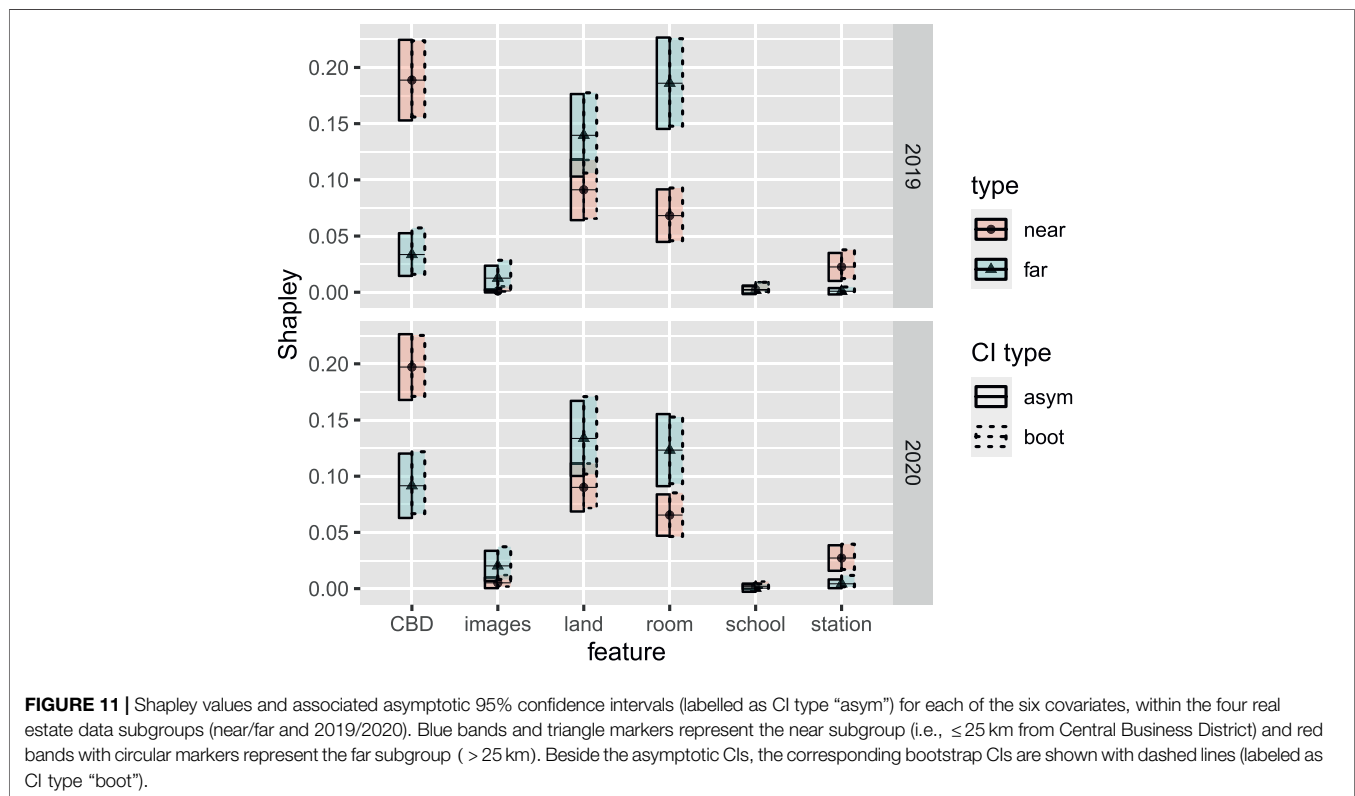
Subgroup	Near (2019)	Far (2019)	Near (2020)	Far (2020)
Sample size	1,203	953	1824	1,130

- For small correlations, in particular $c \leq 0.1$ and sample sizes $10 \leq n \leq 100$, the lower bound of the confidence interval for the coverage of the asymptotic CIs never drops below 0.91.
- For $c = 0$ and $n \geq 15$, the lower bound of the asymptotic CI for coverage never drops below 0.95 in Studies A and B, while in Study C the lower bound is at least 0.88.
- For sample sizes $15 \leq n \leq 50$, the coverage of the asymptotic CI tends to be higher when c is closer to the boundaries of $[0, 1]$, as shown in **Figure 7**.
- The average asymptotic CI width is lower when c is nearer to the boundaries of $[0, 1]$, see **Figure 8**.

We now make some general observations which apply to all three Studies. As sample size increases, the estimated coverage initially increases rapidly, as can be seen, for example, in the left column of **Figure 3**. For small sample sizes between $n = 5$ and $n = 50$, the asymptotic CIs typically outperform bootstrap CIs,

TABLE 3 | Shapley values and asymptotic CIs for the covariates of the real estate data.

Feature	Near (2019)	Far (2019)	Near (2020)	Far (2020)
CBD	0.19 (0.15,0.23)	0.03 (0.02,0.05)	0.20 (0.17,0.23)	0.10 (0.06,0.12)
Images	0.00 (−0.00,0.00)	0.01 (0.00,0.02)	0.06 (0.00,0.01)	0.02 (0.01,0.03)
Land	0.09 (0.06,0.12)	0.14 (0.10,0.18)	0.09 (0.07,0.11)	0.13 (0.10,0.17)
School	0.00 (−0.00,0.00)	0.00 (−0.01,0.01)	0.00 (−0.00,0.00)	0.00 (−0.00,0.00)
Station	0.02 (0.01,0.04)	0.00 (−0.00,0.00)	0.03 (0.02,0.04)	0.00 (0.00,0.01)
Room	0.07 (0.05,0.09)	0.19 (0.15,0.23)	0.07 (0.05,0.08)	0.12 (0.09,0.16)

FIGURE 11 | Shapley values and associated asymptotic 95% confidence intervals (labelled as CI type "asym") for each of the six covariates, within the four real estate data subgroups (near/far and 2019/2020). Blue bands and triangle markers represent the near subgroup (i.e., ≤ 25 km from Central Business District) and red bands with circular markers represent the far subgroup (> 25 km). Beside the asymptotic CIs, the corresponding bootstrap CIs are shown with dashed lines (labeled as CI type "boot").

especially when c lies farther from 0.5; there is a clear drop in coverage as c approaches 0.5, for small samples, as can be seen in **Figure 7**. In many cases, the estimated coverage is above 0.9 for $n \geq 10$. However, empirical coverage does not appear to be an increasing function of sample size in general. On the top row in the left column of **Figure 1**, we observe one example of a clear and extended dip in coverage for n in [100, 1000]. This gives rise to the general observation that the asymptotic intervals have preferable coverage statistics over bootstrap for small samples, but not for a certain range of large samples, depending on c.

We further observe that, for all n, there is a general increase in the average CI width as c approaches 0.5 from either direction, as in **Figure 8**. In all Studies, over all sample sizes and correlations, the bootstrap CI average widths were smaller than the asymptotic CI widths by at most 0.0289, and vice versa by at most 0.0667. In general, the asymptotic intervals display favourable widths, though less so near $c = 0.5$.

4.6. Recommendations for Use
Based on these observations, we recommend using asymptotic CIs over bootstrap CIs under the following conditions:

(i) Computational time is relevant (e.g., estimating a large number of Shapley values).
(ii) The sample size is small (e.g., $n \leq 50$).
(iii) The correlation between explanatory variables and the response variable is expected to be beyond ± 0.2 from 0.5, or when this is where the highest precision is desired.

We note that our observations are made from an incomplete albeit comprehensive set of simulation scenarios. There are of course an infinite number of combinations of simulation cases and thus we cannot guarantee that our observation applies to all possible DGPs.
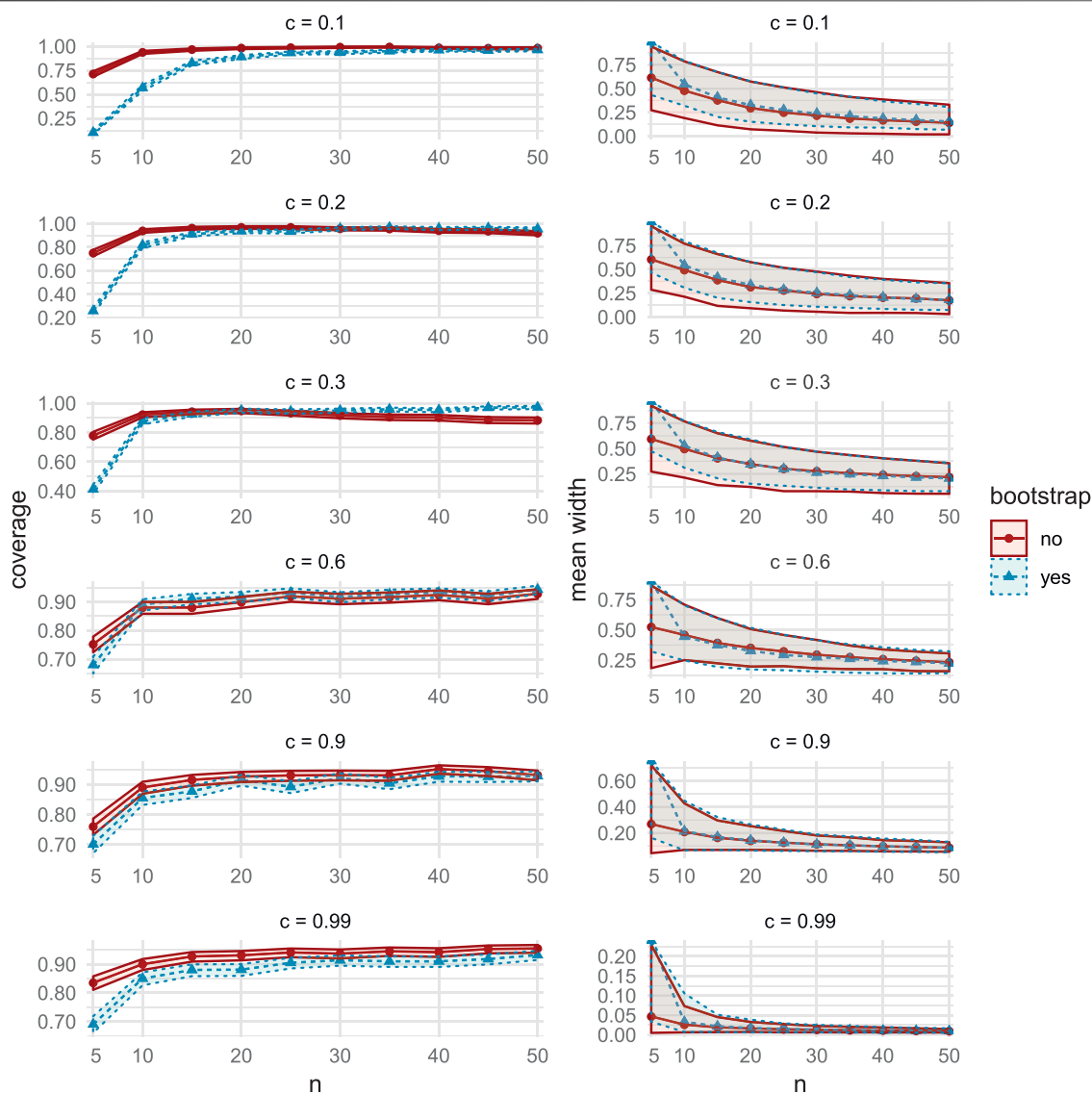
FIGURE 12 | Comparisons of coverage **(left column)** and mean width **(right column)**, between the bootstrap CIs (dashed lines with triangular markers) and the asymptotic CIs (solid lines with circular markers) in MC Study A. Rows represent correlations c, increasing from 0:1 on the top row to 0:99 on the bottom row. The horizontal axes display the sample sizes n = 5;10;50.

5. Application: Real Estate and COVID-19

For an interesting application of our methods, in this section we identify significant changes in the behavior of the local real estate market in Melbourne, Australia, between the period 1 February 2019 to 1 April 2019 and the period 1 February 2020 to 1 April 2020. In 2020, this corresponds to an early period of growing public concern regarding the novel coronavirus COVID-19. We obtain the Shapley decomposition of the coefficient of multiple correlation R^2 between observed house prices and a number of property features. We also find significant differences in behavior between real estate near and far from the Central Business District (CBD), where near is defined to be within 25 km of Melbourne CBD, and far is defined as non-near (see **Figure 9**).

Note that the nature of this investigation is exploratory and expository; our conclusions are not intended to be taken as evidence for the purposes of policy or decision making.

On 1 February the Australian government announced a temporary ban on foreign arrivals from mainland China, and by 1 April a number of social distancing measures were in place. We scraped real estate data from the AUHousePrices website (https://www.auhouseprices.com/), to obtain a data set of 13, 291 (clean) house sales between 1 January and 18 July in each of 2019 and 2020. We then reduced this date range to capture only the spike in sales observed between 1 February and 1 April (see **Figure 10**), giving a remaining sample size of 5,110, which was partitioned into the four subgroups in **Table 2**. By capturing this
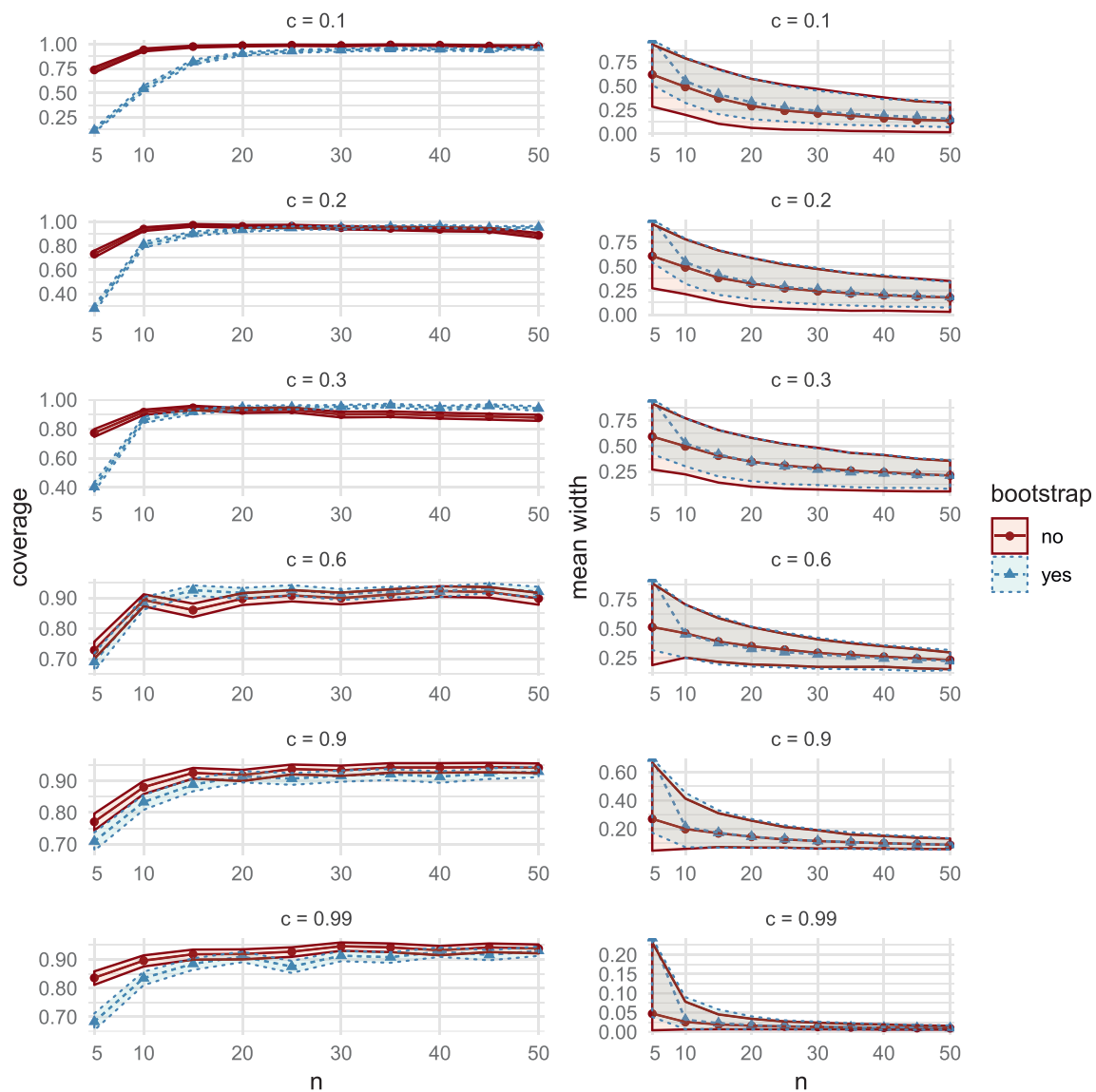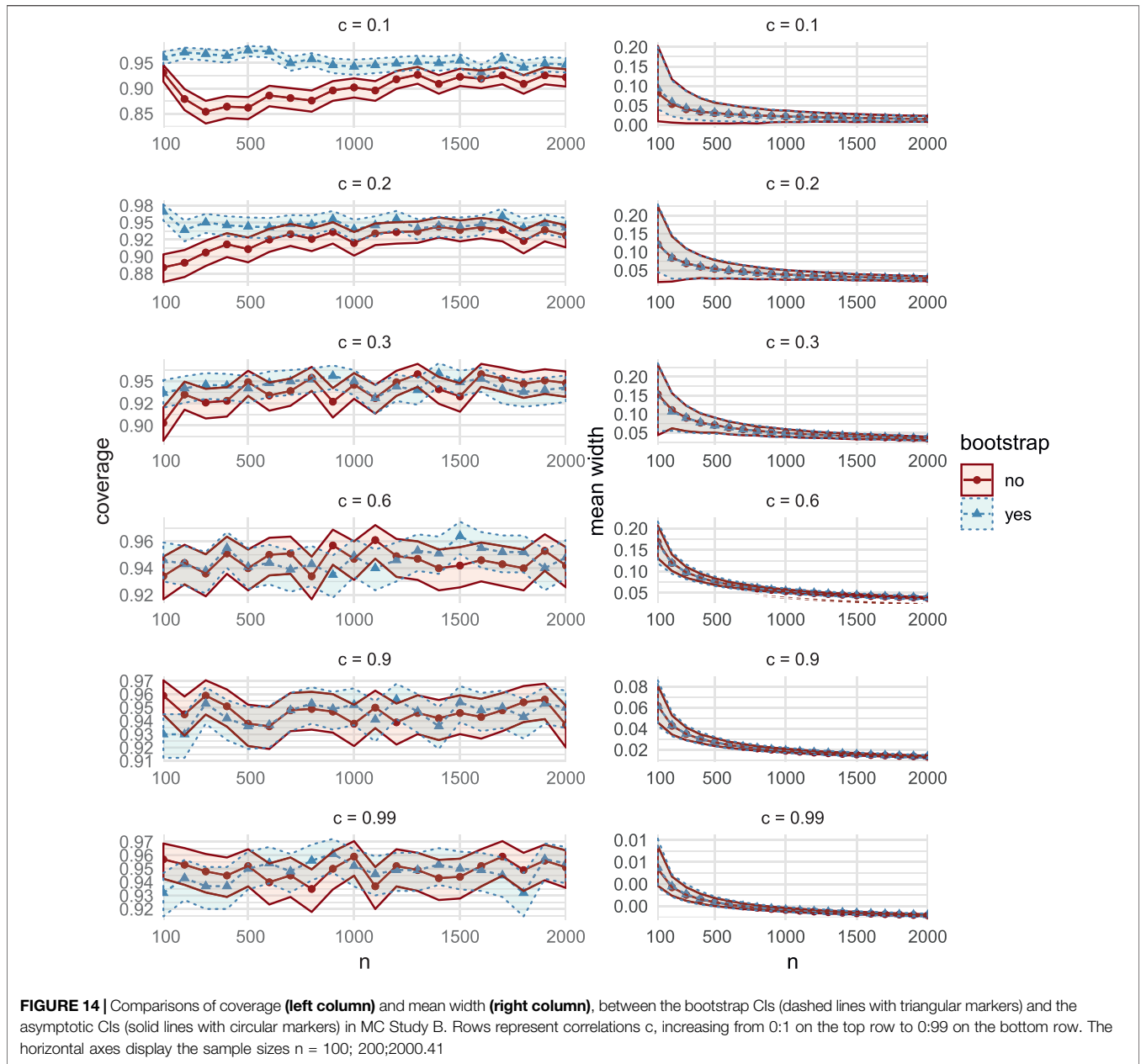
FIGURE 13 | Comparisons of coverage **(left column)** and mean width **(right column)**, between the bootstrap CIs (dashed lines with triangular markers) and the asymptotic CIs (solid lines with circular markers) in MC Study B. Rows represent correlations c, increasing from 0:1 on the top row to 0:99 on the bottom row. The horizontal axes display the sample sizes n = 5; 10;50.

spike in sales, we hope to detect a significant change in the Shapley values between the same period in 2019 and 2020. Within each of the four subgroups we perform a Yeo-Johnson transformation [29] to reduce any violation of the assumption of joint pseudo-ellipticity.

We decompose R^2 among the covariates: distance to CBD (CBD); images used in advertisement (images); property land size (land); distance to nearest school (school); distance to nearest station (station); and number of bedrooms + bathrooms + car ports (room); along with the response variable, house sale price (price). We expect the room covariate to act as a proxy for house size. Thus we decompose R^2 for the linear model,

$$(\text{price}) = \beta_0 + \boldsymbol{\beta}\,(\text{CBD, images, land, school, station, room})^T,$$
$$\beta_0 \in \mathbb{R}, \ \boldsymbol{\beta} \in \mathbb{R}^6. \tag{14}$$

Fitted to each of the four subgroups, we obtain $R^2 = 0.37$ for model **Eq. 14**, for each subgroup except for the near (2020) subgroup, for which $R^2 = 0.39$. The resulting Shapley values and 95% confidence intervals are listed in **Table 3**, and shown graphically in **Figure 11**. Also in **Figure 11**, the corresponding bootstrap CIs are given for completeness, and are almost identical to the asymptotic CIs. From these results we make the following observations regarding attributions of the total variability in house prices explained by model **Eq. 14**:

FIGURE 14 | Comparisons of coverage **(left column)** and mean width **(right column)**, between the bootstrap CIs (dashed lines with triangular markers) and the asymptotic CIs (solid lines with circular markers) in MC Study B. Rows represent correlations c, increasing from 0:1 on the top row to 0:99 on the bottom row. The horizontal axes display the sample sizes n = 100; 200;2000.41

(i) In both 2019 and 2020, the attribution for distance to CBD was significantly higher for house sales near to the CBD, compared to house sales farther from the CBD. Correspondingly, the attribution for roominess was significantly lower for house sales near to the CBD.

(ii) Among sales that were near to the CBD, distances to the CBD received significantly greater attribution than both land size and roominess in 2019. Unlike the case for sales that were far from the CBD, these differences remained significant in 2020.

(iii) Distances to stations and schools, as well as images used in advertising, have apparently had little overall impact. In all four subgroups, the attribution for distance to the nearest school is not significantly different from 0. However, distance to a station does receive significantly more attribution among houses that are near to the CBD, compared to those farther away.

(iv) Interestingly, while not a significant result, the number of images used in advertising did appear to receive greater attribution among house sales that were far from the CBD, compared to those near to it.

(v) Among sales that were far from the CBD, land size and roominess both received significantly more attribution than distance to the CBD, in 2019. However, this difference vanished in 2020, with distances to the CBD apparently gaining more attribution, while roominess and land size apparently lost some attribution, in a relative leveling out of these three covariates.
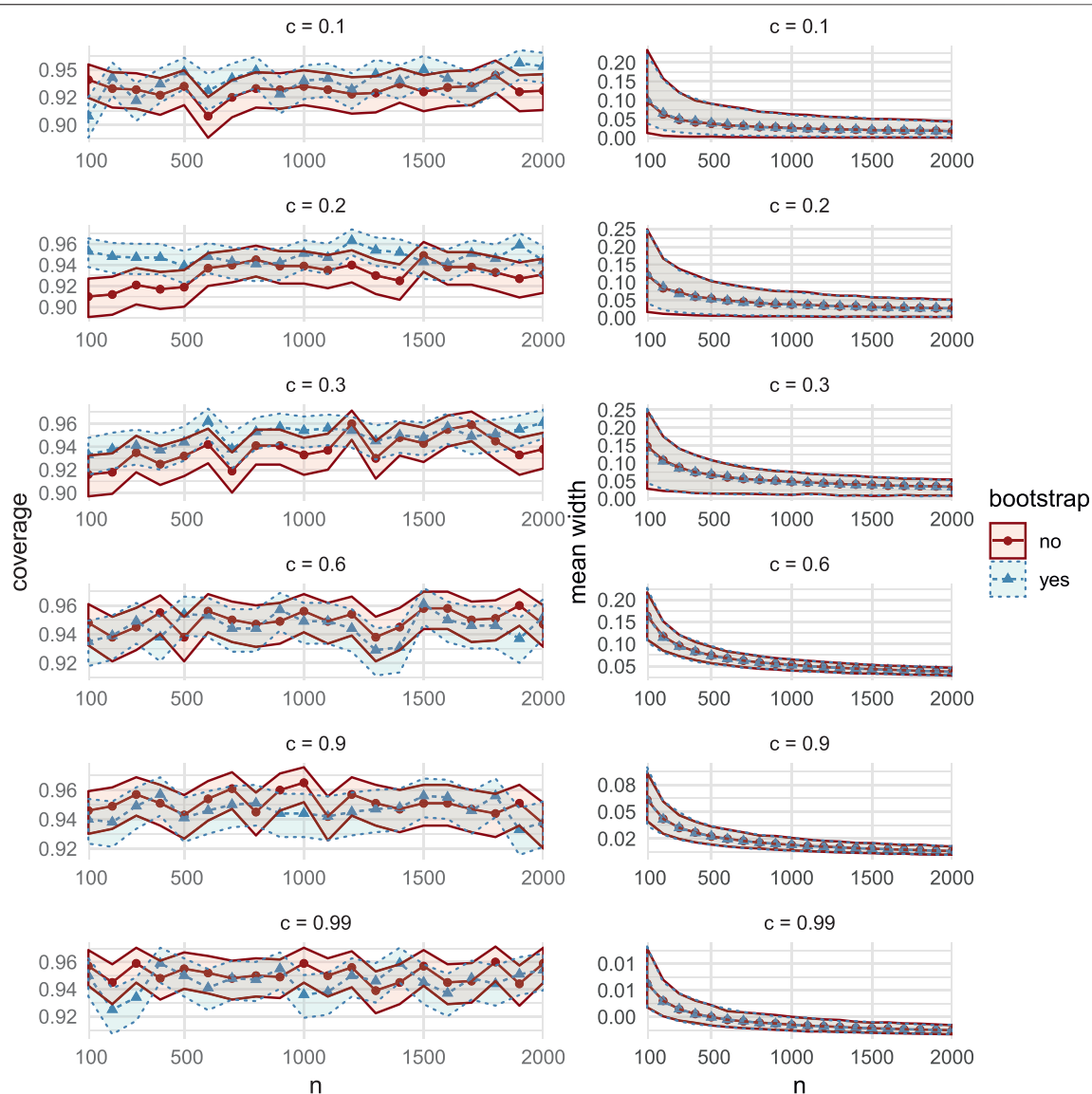
FIGURE 15 | Comparisons of coverage **(left column)** and mean width **(right column)**, between the bootstrap CIs (dashed lines with triangular markers) and the asymptotic CIs (solid lines with circular markers) in MC Study C. Rows represent correlations c, increasing from 0:1 on the top row to 0:99 on the bottom row. The horizontal axes display the sample sizes n = 100; 200; 2000.

Item (i) is perhaps unsurprizing: distances were less relevant far from the city, where price variability was influenced more by roominess and land size. Indeed, we can assume we are less likely to find small and expensive houses far from the CBD. However, the authors find Item (v) interesting: near the city, the behavior did not change significantly during the 2020 period. However, far from the city, the behavior did change significantly, moving toward the near-city behavior. Distance to the city became more important for explaining variability in price, while land size and roominess both became less important, compared with their 2019 estimates. Our initial guess at an explanation was that near-city buyers, with near-city preferences, were temporarily motivated by urgency to buy

farther out. However, according to **Table 2**, the observed ratio of near-city buyers to non-near buyers actually increased in this period, from 1.26 in 2019 to 1.61 in 2020. We will not take this expository analysis further here, but we hope that the interested reader is motivated to take it further, and to this end we have made the data, as well as R and Julia scripts, available at https://github.com/ex2o/shapley_confidence.

6. DISCUSSION

In this work, we focus on regression Shapley values as a means for attributing goodness of fit. There has also been much recent

interest in the machine learning literature in using Shapley values to attribute predictions. These approaches, such as SHAP [30], TreeExplainer [31], and Quantitative Input Influence [32] are focused on providing local explanations that depend on the behavior of a fitted model. For more on the differences between prediction, estimation and attribution, see the recent exposition of Ref. 33.

Research into the attribution of goodness of fit using the Shapley value, predates research into attributing predictions. As described in **Section 2**, regression Shapley values were first introduced in Ref. 10. Then, Ref. 1 advocated for decompositions among exogenously grouped sets of explanatory variables using the both the Shapley and Owen values, where the Owen values are a generalization that permits group attributions.

Our work is the first to determine the asymptotic distribution of the regression Shapley values. In **Section 3**, we show that under an elliptical (or pseudo-elliptical) joint distribution assumption, the Shapley value decomposition of $R^2(\mathcal{Z}_n)$ is asymptotically normal. This involved correcting a number of errors that appeared in the expositions of Lemmas 2 and 3 in prior sources, and stating those results in a notation conforming to the language of cooperative games and set-valued functions. Our result regarding the asymptotic distribution leads immediately to Shapley value confidence intervals and hypothesis tests.

In **Section 4**, we use Monte Carlo simulations to examine the coverage and width statistics of these asymptotic CIs. We compare the coverage and width statistics for different data generating processes, correlations and sample sizes. In **Section 4.5** we provide recommendations for when asymptotic CIs should be used rather than bootstrap CIs. We find that the asymptotic CIs have estimated coverage probabilities of at least 0.85 across all studies, are preferable over the bootstrap CIs for small sample sizes ($n \leq 50$), and are often (although not always) favourable for large sample sizes. We discuss the generality of our findings that asymptotic CIs are computationally much more efficient than bootstrap CIs.

In **Section 5**, we derive asymptotic CIs to Melbourne house prices during a period of altered consumer behavior in the initial stages of the arrival of COVID-19 in Australia. Using the CIs, we identify significant changes in model behavior between 2019 and 2020, and attribute these changes among features, highlighting an interesting direction of future research into the period. Implementations of our methods as well as data sets are openly available for use at https://github.com/ex2o/shapley_confidence.

We aim to use our developments to derive the asymptotic distributions of variance inflation factors and their generalizations [34], as well as the closely related Owen values decomposition of the coefficient of determination for exogenously grouped sets of features [1]. We will also investigate calculating Shapley values capable of uncovering non-linear dependencies and their confidence intervals in future work.

AUTHOR'S NOTE

This manuscript has been released as a pre-print on arXiv [35].

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

DF Provided Monte Carlo simulations, theoretical contributions, real estate data scraping and analysis, figures, computational benchmarks and general writing. IS Provided real estate data analysis, theoretical contributions, general writing and proofreading. HN Provided theoretical proofs, theoretical contributions, general writing and proofreading.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fams.2020.587199/full#supplementary-material

REFERENCES

1. Huettner F, Sunder M. Decomposing axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values. *Electron J Statist.* (2012). 6:1239–50. doi:10.1214/12-ejs710

2. Efron B. Computer-intensive methods in statistical regression. *SIAM Rev.* (1988). 30:421–49. doi:10.1137/1030093

3. Baglivo JA. *Mathematica laboratories for mathematical statistics: emphasizing simulation and computer intensive methods*. Philadelphia, PA: SIAM (2005).

4. Steiger JH, Browne MW. The comparison of interdependent correlations between optimal linear composites. *Psychometrika* (1984). 49:11–24. doi:10.1007/bf02294202

5. Hedges LV, Olkin I. Joint distributions of some indices based on correlation coefficients. In: S Karlin, T Amemiya, LA Goodman, editors *Studies in econometrics, time series, and multivariate analysis*. Cambridge, UK: Academic Press (1983).

6. Olkin I, Finn JD. Correlations redux. *Psychol Bull* (1995). 118:155–64. doi:10.1037/0033-2909.118.1.155

7. Seber GAF, Lee AJ. *Linear regression analysis*. Hoboken, NJ: Wiley (2003).

8. Greene WH. *Econometric analysis*. Upper Saddle River, NJ: Prentice-Hall. (2007).

9. Cowden DJ. The multiple-partial correlation coefficient. *J Am Stat Assoc.* (1952). 47:442–56. doi:10.1080/01621459.1952.10501183

10. Lipovetsky S, Conklin M. Analysis of regression in game theory approach. *IED Stoch Models Bus Ind Appl.* 17, 319–30. doi:10.1002/asmb.446

11. Israeli O (2007). A Shapley-based decomposition of the R-square of a linear regression. *J Econ Inequal.* 5, 199–212.

12. Shapley LS. A value for n-person games. In: HW Kuhn AW Tucker, editors *Contributions to the theory of games*. Princeton, NJ: Princeton University Press (1953).

13. Gromping U. Relative importance for linear regression R: the package relaimpo. *J Stat Softw.* (2006). 17, 1–27.

14. Grömping U. Estimators of relative importance in linear regression based on variance decomposition. *Am Stat.* (2007). 61, 139–47. doi:10.1198/000313007x188252

15. Young HP. Monotonic solutions of cooperative games. *Int J Game Theory* (1985). 14:65–72. doi:10.1007/bf01769885

16. Feltkamp V. Alternative axiomatic characterizations of the shapley and banzhaf values. *Int J Game Theory* (1995). 24(2):179–86. doi:10.1007/bf01240041

17. Kruskal W. Relative importance by averaging over orderings. *Am Am Stat.* (1987). 41, 6–10. doi:10.2307/2684310

18. Genizi A. Decomposition of R^2 in multiple regression with correlated regressors. *Stat Sin.* (1993). 13:407–20.

19. Mardia KV (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57:519–30. doi:10.1093/biomet/57.3.519

20. Fang KT, Kotz S, Ng KW. *Symmetric multivariate and related distributions.* London, UK: Chapman & Hall (1990).

21. Yuan K-H, Bentler PM. On normal theory and associated test statistics in covariance structure analysis under two classes of nonnormal distributions. *Stat Sin.* (1999). 9:831–53.

22. Yuan K-H, Bentler PM. Inferences on correlation coefficients in some classes of nonnormal distributions. *J Multivar Anal.* (2000). 72:230–48. doi:10.1006/jmva.1999.1858

23. Hedges LV, Olkin I. Joint distribution of some indices based on correlation coefficients (1983). Technical Report MCS 81-04262, Stanford University.

24. van der Vaart A. *Asymptotic statistics.* Cambridge, UK: Cambridge University Press (1998).

25. Seber GAF. *A matrix Handbook for statisticians.* New York, NY: Wiley (2008).

26. Hart S, Mas-Colell A. The potential of the shapley value. *Shapley Value* (1988). 14:127–37.

27. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* (1934). 26(4):404–13. doi:10.1093/biomet/26.4.404

28. Fujikoshi Y, Ulyanov VV, Shimizu R. *Multivariate statistics: high-dimensional and large-sample approximations.* Vol. 760. Hoboken, NJ: John Wiley & Sons (2011).

29. Yeo I-K, Johnson RA. A new family of power transformations to improve normality or symmetry. *Biometrika* (2000). 87(4), 954–59. doi:10.1093/biomet/87.4.954

30. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, editors *Advances in neural information processing systems.* Vol. 30, London, UK: Curran Associates, Inc (2017). p. 4765–74.

31. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable ai for trees. *Nat Mach Intell.* (2020). 2(1):56–67. doi:10.1038/s42256-019-0138-9

32. Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In: 2016 IEEE symposium on security and privacy (SP); 22–26 May 2016; San Jose, CA. IEEE. (2016). p. 598–617.

33. Efron B. Prediction, estimation, and attribution. *J Am Stat Assoc.* (2020). 115(530):636–55. doi:10.1080/01621459.2020.1762613

34. Fox J, Monette G. Generalized collinearity diagnostics. *J Am Stat Assoc.* (1992). 87(417):178–83. doi:10.1080/01621459.1992.10475190

35. Fryer D, Strumke I, Nguyen H. *Shapley value confidence intervals for variable selection in regression models.* arXiv preprint arXiv:2001.09593 (2020).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.