universität wien

# MAGISTERARBEIT / MASTER'S THESIS

Titel der Magisterarbeit / Title of the Master's Thesis

„Mathematical aspects of applying conformal prediction to regression models“


verfasst von / submitted by

Dmytro Rzhemovskyi


angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magister der Sozial- und Wirtschaftswissenschaften
(Mag. rer. soc. oec.)

Wien, 2023 / Vienna, 2023

| | |
|---|---|
| Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet: | UA 066 951 |
| Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet: | Statistik |
| Betreut von / Supervisor: | Univ.-Prof. Mag. Dr. Hannes Leeb |

**Abstract**

Conformal prediction (CP) is a method of constructing prediction intervals presented 20 years ago by Vovk and Gammerman. In this thesis we provide an overview of CP and consider important properties of the method such as strong validity and optimality as well as its applications to the regression problem under different settings. We focus on providing rigorous proofs for the results presented in the paper "On-line predictive linear regression" and consider in detail all the necessary additional concepts.

**German Abstract**

Conformal prediction (CP) ist eine Methode zur Konstruktion von Vorhersageintervallen, die vor 20 Jahren von Vovk und Gammerman eingeführt wurde. In dieser Arbeit geben wir einen Überblick über CP und betrachten wichtige Eigenschaften dieser Methode wie starke Validität und Optimalität sowie ihre Anwendungen auf das Regressionsproblem unter verschiedenen Voraussetzungen. Der Fokus liegt auf rigorosen Beweisen der Ergebnisse, die in der Arbeit "On-line predictive linear regression" präsentiert werden, wobei alle zusätzlichen relevanten Konzepte einbezogen werden.

# Contents

# 1 Introduction

Consider the following classical regression problem in statistics and machine learning:

Given the previous observations $(x_1, y_1), \dots, (x_{n-1}, y_{n-1}) \in \mathbb{R}^K \times \mathbb{R}$, generated independently from some unknown distribution $P$, and a label $x_n$ of a new observation, we want to construct a prediction for the target value $y_n$.

Usually, prediction for the variable $y_n$ is understood as "point prediction" $\hat{y}_n \in \mathbb{R}$. For this interpretation, there is a well-developed theory and a wide range of methods: linear regression, neural networks, decision trees, etc.

On the other hand, we would like to have as a prediction not only a single point, but a set of possible values together with the probability $1 - \varepsilon \in [0, 1]$ that our prediction set will contain the target variable $y_n$, i.e.,

$$\Gamma_n^\varepsilon = \Gamma_n^\varepsilon\big((x_1, y_1), \dots, (x_{n-1}, y_{n-1}), x_n\big) \subset \mathbb{R}$$
$$\text{such that } \mathbb{P}\big(y_n \in \Gamma_n^\varepsilon\big) = 1 - \varepsilon.$$

20 years ago, in their book "Algorithmic learning in a random world" [1] and their previous series of works, Vovk and Gammerman proposed a universal method of constructing valid prediction sets: conformal prediction. The idea behind the method is to predefine a function that measures how different a new observation $(x_n, y_n)$ from the previous observations $(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$ is. Then it is natural to select only hypothetical values $y \in \mathbb{R}$ of the target variable $y_n$ such that $(x_n, y)$ will be not too unlike to the previous data points.

One of the advantages of the method is that it allows to construct valid prediction intervals under quite a weak assumption of exchangeability of the data: the joint distribution of the data random variables is invariant under permutations of the data points (see the details in the next section).

Throughout the thesis, we will work under the so-called on-line protocol. Unlike the classical "batch" model, where the number of data points is fixed, we assume that data points are generated sequentially. Therefore, we should also explore the properties of a sequence of prediction regions $\Gamma_n^\varepsilon$, $n \geq 2$. For example, as we will see in Subsection 3.2, on each step $n$ the coverage of target variable $y_n$ by the conformal interval $\Gamma_n$ is independent from the coverage on other steps. This property is referred to as strong validity.

Another important property of the method is that some classical prediction intervals can be obtained as conformal intervals. For example, in case of normally distributed random variables $x_1, \dots, x_n$ with unknown mean and variance, to predict $x_n$, we can use the Student's prediction interval

$$C_n^\varepsilon = C_n^\varepsilon(x_1, \dots, x_{n-1}) = \Big[\overline{x}_{n-1} \pm t_{n-2}^{1-\varepsilon/2} \hat{\sigma}_{n-1} \sqrt{\frac{n}{n-1}}\Big].$$

In Subsection 4.2 we show that this prediction interval can be also constructed as a conformal interval, this shows that strong validity holds for the Student's prediction

interval. It should be mentioned that the latter fact was not known before the introduction of conformal prediction.

As we discuss in Subsection 3.3, conformal intervals are in some sense optimal confidence intervals, i.e. under certain conditions, for every confidence interval, one may construct a conformal interval that is not worse (not larger) and has the same probability of coverage.

In this thesis, we provide an overview and detailed proofs of the results from the paper "On-line predictive linear regression" [2] together with the supplementary theory.

# 2 Illustrative example

## 2.1 Conformal intervals for the exchangeability model

To discuss the basic ideas and concepts of conformal prediction, we start with the case of the exchangeability model. A short overview of the concept of exchangeable random variables can be found in Subsection 2.2.

Let $\mathcal{P}$ be a set of probability measures such that under every $\mathbb{P} \in \mathcal{P}$ the joint distribution of random variables $Y_1, \dots, Y_n$ is continuous and exchangeable.

Recall that our goal is to construct a prediction interval for the random variable $Y_n$ using the sequence of "previous observations" $Y_1, \dots, Y_{n-1}$, i.e., for every significance level $\varepsilon \in (0,1)$, we want to build a set $\Gamma_n^\varepsilon = \Gamma_n^\varepsilon(Y_1, \dots, Y_{n-1}) \subset \mathbb{R}$ such that

$$\mathbb{P}\big(Y_n \in \Gamma_n^\varepsilon\big) = \varepsilon$$

under any distribution $\mathbb{P} \in \mathcal{P}$.

**Definition 2.1.** Let $\mathcal{P}$ be a statistical model (the set of probability measures, one of which we believe generates the data) and let $Y_1, \dots, Y_n$ be a sequence of random variables. We say that statistic $S_n = S_n(Y_1, \dots, Y_n)$ is sufficient for the model $\mathcal{P}$ (or for the sequence $Y_1, \dots, Y_n$) if the conditional joint distribution of $Y_1, \dots, Y_n$ given $S_n$ under $\mathbb{P}$ is the same for all $\mathbb{P} \in \mathcal{P}$.

As we can see from Theorem 2.3, the vector of order statistics is a sufficient statistic in the continuous exchangeability model. Now applying conditioning on the order statistic, we get a known and fixed distribution of the sequence $Y_1, \dots, Y_n$, which we will use for our further constructions.

For convenience, we introduce some notation:

- $Y_{(\cdot),n}$ is a vector of order statistics in the sample $Y_1, \dots, Y_n$.
- $Y_{(\cdot),n,k}$ is a vector of order statistics in the sample $Y_1, \dots, Y_n$ excluding $Y_k$.
- $\overline{Y}_n$ and $\overline{Y}_{n,k}$ are means of the vectors $Y_{(\cdot),n}$ and $Y_{(\cdot),n,k}$ respectively .

Let us briefly describe the idea behind conformal prediction. For every hypothetical value $y \in \mathbb{R}$ of the random variable $Y_n$ we will decide whether to include it in the prediction interval using the following criteria: if $y$ is too different (nonconformal) from the previous observations $Y_1, \dots, Y_{n-1}$, then we do not select it.

To measure the difference between a hypothetical value and previous observations we will use a specific function called nonconformity measure. In this chapter, we consider the function

$$A\big((y_1, \dots, y_{n-1}), y_n\big) = |\overline{y}_{n-1} - y_n|. \tag{1}$$

Values of nonconformity functions we call nonconformity scores.

Now given the value of $A\big((Y_1, \dots, Y_{n-1}), y\big)$, we need to understand whether the difference between $Y_1, \dots, Y_{n-1}$ and $y$ is not critical. We can define it by comparing $A\big((Y_1, \dots, Y_{n-1}), y\big)$ with other nonconformity scores in the sample given by

$$A^{(i)} := A\big(Y_{(\cdot),n,i}, Y_i\big),\ i = 1, \dots, n \tag{2}$$

where we assume that $Y_n = y$.

Note that the nonconformity measure (1) does not depend on the order of the previous values $y_1, \dots, y_{n-1}$ (i.e., their permutations will produce the same nonconformity score) and hence we get no more than $n$ distinct nonconformity scores. Together with the assumption about the continuity of the joint distribution, we obtain that almost surely the values of $Y_1, \dots, Y_n$ are distinct and therefore we have exactly $n$ nonconformity scores.

The critical value mentioned above can be calculated by a standard statistical approach using $p$-values. Roughly speaking, the $p$-value is the probability to get a more extreme result than what we observe. In our case, since the conditional distribution of $Y_1, \dots, Y_n$ given $Y_{(\cdot),n}$ is independent of the probability measure $\mathbb{P} \in \mathcal{P}$, we can consider a $p$-value of the observed nonconformity score in the form

$$\begin{aligned} p_n &= p_n\big((y_1, \dots, y_{n-1}), y_n\big) \\ &= \mathbb{P}\big(A((Y'_1, \dots, Y'_{n-1}), Y'_n) \geq A((y_1, \dots, y_{n-1}), y_n) \mid Y'_{(\cdot),n} = y_{(\cdot),n}\big) \end{aligned}$$

where we assume that $(Y'_1, \dots, Y'_n)$ is an iid copy of $(Y_1, \dots, Y_n)$.

From Theorem 2.3, we know that the conditional distribution of $Y_1, \dots, Y_n$ given $Y_{(\cdot),n}$ is uniform over the set of permutations of $Y_{(\cdot),n}$. Therefore the conditional distribution of the nonconformity measure is also uniform over the set $\big\{\, A^{(i)} \mid i = 1, \dots, n \,\big\}$ and

$$p_n\big((y_1, \dots, y_{n-1}), y_n\big) = \frac{\#\big\{\, i = 1, \dots, n \mid A^{(i)} \geq A^{(n)} \,\big\}}{n} \tag{3}$$

for distinct values $y_1, \dots, y_n \in \mathbb{R}$.

Now to construct conformal intervals, we will use the following principle: It is natural to include in the prediction interval such hypothetical values $y \in \mathbb{R}$ that have not too large "relative nonconformity" defined by the corresponding $p$-value (3). Hence for the confidence rate $\varepsilon \in (0, 1)$, we get

$$\Gamma^\varepsilon_n(y_1, \dots, y_{n-1}) = \big\{\, y \in \mathbb{R} \mid p_n((y_1, \dots, y_{n-1}), y) \geq \varepsilon \,\big\}. \tag{4}$$

The following result guarantees the coverage of $Y_n$ by the prediction interval defined in (4) with the probability at least $1 - \varepsilon$.

**Theorem 2.1.** Let $Y_1, \dots, Y_n$ be a sequence of jointly continuous and exchangeable random variables. Then for every $\varepsilon \in (0, 1)$ and conformal interval $\Gamma^\varepsilon_n$ defined in (4) it holds that

$$\mathbb{P}\big(Y_n \in \Gamma^\varepsilon_n\big) \geq 1 - \varepsilon.$$

*Proof.* First notice that conditionally on $Y_{(\cdot),n}$, $p$-value $p_n = p_n\big((Y_1, \dots, Y_{n-1}), Y_n\big)$ is uniformly distributed over the set $\{\frac{1}{n}, \dots, \frac{n-1}{n}, 1\}$ almost surely, because every value of the nonconformity score $A^{(n)}$ corresponds to some unique value of $p_n$:

$$\frac{\#\big\{\, i = 1, \dots, n \mid A^{(i)} \geq A^{(n)} \,\big\}}{n} \in \Big\{\frac{1}{n}, \dots, \frac{n-1}{n}, 1\Big\}.$$

Since conditionally on $Y_{(\cdot),n}$ the nonconformity measure $A$ is uniformly distributed, we get that $p_n$ is also uniformly distributed.

Now we have that

$$\begin{aligned}\mathbb{P}\big(Y_n \in \Gamma_n^\varepsilon \mid Y_{(\cdot),n}\big) &= \mathbb{P}\big(p_n((Y_1,\dots,Y_{n-1}),Y_n) \geq \varepsilon \mid Y_{(\cdot),n}\big) \\ &= \mathbb{P}\Big(p_n \geq \frac{\lceil n\varepsilon \rceil}{n} \,\Big|\, Y_{(\cdot),n}\Big) \geq 1-\varepsilon\end{aligned}$$

which implies that $\mathbb{P}(Y_n \in \Gamma_n^\varepsilon) = \mathbb{E}\mathbb{P}(Y_n \in \Gamma_n^\varepsilon \mid Y_{(\cdot),n}) \geq 1-\varepsilon$.

□

Now, if we want that prediction interval to cover $Y_n$ with probability exactly equal to $1-\varepsilon$, we can use the procedure of randomization. Additional randomization allows us to obtain a continuous $p$-value from a discrete one in the following way.

For $t_n \in [0,1]$, define a randomized $p$-value as

$$p_n = p_n((y_1,\dots,y_{n-1}),y_n,t_n) = \frac{\#\big\{\, i=1,\dots,n \mid A^{(i)} > A^{(n)} \,\big\}}{n} + \frac{t_n}{n}.$$

We call the function $p_n$ randomized, since later we will use a $U[0,1]$-distributed random variable instead of $t_n$.

Define the corresponding conformal interval as

$$\Gamma_n^\varepsilon(t_n) = \big\{\, y \in \mathbb{R} \mid p_n\big((y_1,\dots,y_{n-1}),y,t_n\big) \geq \varepsilon \,\big\}. \tag{5}$$

**Theorem 2.2.** If $\tau_n$ is a $U([0,1])$-distributed random variable, independent of the sequence of exchangeable and jointly continuous random variables $Y_1,\dots,Y_n$, then for any $\varepsilon \in (0,1)$ and the corresponding conformal interval $\Gamma_n^\varepsilon(\tau_n)$ defined in (5) it holds that

$$\mathbb{P}\big(Y_n \in \Gamma_n^\varepsilon(\tau_n)\big) = 1-\varepsilon.$$

*Proof.* First we note that conditionally on $Y_{(\cdot),n}$, the randomized $p$-value $p_n(\tau_n)$ is $U([0,1])$-distributed, since, conditionally on $Y_{(\cdot),n}$, the random variable

$$R_n = \frac{\#\big\{\, i=1,\dots,n \mid A^{(i)} > A^{(n)} \,\big\}}{n}$$

is uniformly distributed on the set $\big\{\, 0, \frac{1}{n}, \dots, \frac{n-1}{n} \,\big\}$ and $\tau_n/n$ is $U([0,\frac{1}{n}])$-distributed. Since both random variables are independent given $Y_{(\cdot),n}$, we get for $\varepsilon \in (k/n, (k+1)/n]$ that

$$\begin{aligned}&\mathbb{E}\big(\mathbb{1}_{\{p_n(\tau_n)<\varepsilon\}} \mid Y_{(\cdot),n}\big) \\ &\quad= \mathbb{E}\big(\mathbb{1}_{\{R_n<k/n\}} + \mathbb{1}_{\{R_n=k/n\}}\mathbb{1}_{\{\tau_n/n\in[0,\varepsilon-k/n)\}} \mid Y_{(\cdot),n}\big) \\ &\quad= \mathbb{E}\big(\mathbb{1}_{\{R_n<k/n\}} \mid Y_{(\cdot),n}\big) + \mathbb{E}\big(\mathbb{1}_{\{R_n=k/n\}} \mid Y_{(\cdot),n}\big)\,\mathbb{E}\big(\mathbb{1}_{\{\tau_n/n\in[0,\varepsilon-k/n)\}} \mid Y_{(\cdot),n}\big) \\ &\quad= \mathbb{E}\big(\mathbb{1}_{\{R_n<k/n\}} \mid Y_{(\cdot),n}\big) + \mathbb{E}\big(\mathbb{1}_{\{R_n=k/n\}} \mid Y_{(\cdot),n}\big)\,\mathbb{E}\big(\mathbb{1}_{\{\tau_n/n\in[0,\varepsilon-k/n)\}}\big) \\ &\quad= \frac{k}{n} + \frac{1}{n}\,\mathbb{P}\big(\tau_n \in [0,n\varepsilon-k)\big) = \frac{k}{n} + \frac{1}{n}(n\varepsilon-k) = \varepsilon.\end{aligned}$$

Now, to compute the probability of covering the random variable $Y_n$ by the prediction interval $\Gamma_n^\varepsilon(\tau_n)$, we can integrate the corresponding conditional probability given $Y_{(\cdot),n}$

$$\mathbb{P}\big(Y_n \in \Gamma_n^\varepsilon(\tau_n)\big) = \mathbb{E}\mathbb{P}\big(Y_n \in \Gamma_n^\varepsilon(\tau_n) \mid Y_{(\cdot),n}\big) = \mathbb{E}\mathbb{P}\big(p_n(\tau_n) \geq \varepsilon \mid Y_{(\cdot),n}\big) = 1 - \varepsilon.$$

□

## 2.2 Exchangeability

**Definition 2.2.** A finite sequence of random variables $Y_1, \ldots, Y_n$ is said to be exchangeable if for every $B^n \in \mathcal{B}(\mathbb{R}^n)$ and every permutation $\pi \in \mathcal{S}_n$ it holds that

$$\mathbb{P}\big((Y_1, \ldots, Y_n) \in B^n\big) = \mathbb{P}\big((Y_{\pi(1)}, \ldots, Y_{\pi(n)}) \in B^n\big).$$

An infinite sequence of random variables $Y_1, Y_2, \ldots$ is said to be exchangeable if for every $n \geq 2$ the finite sequence $Y_1, \ldots, Y_n$ is exchangeable.

*Remark.* By taking $B^n = B \times \mathbb{R} \times \cdots \times \mathbb{R}$, where $B \in \mathcal{B}(\mathbb{R})$, we have that

$$\mathbb{P}(Y_1 \in B) = \mathbb{P}((Y_1, \ldots, Y_n) \in B^n) = \mathbb{P}((Y_{\pi(1)}, \ldots, Y_{\pi(n)}) \in B^n) = \mathbb{P}(Y_{\pi(1)} \in B)$$

what implies that random variables in an exchangeable sequence are identically distributed.

*Remark.* If random variables $Y_1, \ldots, Y_n$ are independent and identically distributed, then for any $B^n = B_1 \times \cdots \times B_n \in \mathcal{B}(\mathbb{R}^n)$, where $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$, we have

$$\mathbb{P}((Y_1, \ldots, Y_n) \in B^n) = \prod_{i=1}^{n} \mathbb{P}(Y \in B_i) = \mathbb{P}\big((Y_{\pi(1)}, \ldots, Y_{\pi(n)}) \in B^n\big),$$

which implies that any sequence of iid random variables is exchangeable.

**Theorem 2.3.** Let $Y = (Y_1, \ldots, Y_n)$ be a sequence of exchangeable random variables with a continuous joint distribution. Denote as $Y_{(\cdot),n}$ the vector of order statistics $\big(Y_{(1)}, \ldots, Y_{(n)}\big)$. Then for every permutation $\pi \in \mathcal{S}_n$, we have that

$$\mathbb{P}\big(Y_{\pi(1)} < \cdots < Y_{\pi(n)} \mid Y_{(\cdot),n}\big) = \frac{1}{n!} \text{ a.s.}$$

*Proof.* Let $O = \big\{\, (y_1, \ldots, y_n) \in \mathbb{R}^n \mid y_1 < \cdots < y_n \,\big\}$. For a permutation $\pi \in \mathcal{S}_n$, we write

$$\pi(y_1, \ldots, y_n) = (y_{\pi(1)}, \ldots, y_{\pi(n)})$$
$$\pi(B) = \big\{\, \pi(y) \mid y \in B \,\big\} \text{ for } B \subset \mathbb{R}^n.$$

According to the definition of conditional probability, we need to prove that for every $B \in \mathcal{B}(\mathbb{R}^n)$, it holds that

$$\mathbb{P}\big((Y_{\pi(1)}, \ldots, Y_{\pi(n)}) \in O, Y_{(\cdot),n} \in B\big) = \mathbb{E}\Big(\frac{1}{n!}\, \mathbb{1}_{\{Y_{(\cdot),n} \in B\}}\Big).$$

Note that intuitively, due to exchangeability, every value of the vector of order statistics $Y_{(\cdot),n}$ corresponds almost surely to its $n!$ equally probable permutations. Formally, we have

$$\begin{aligned}\mathbb{P}\big(Y_{(\cdot),n} \in B\big) &= \mathbb{P}\big(Y_{(\cdot),n} \in B \cap O\big) \\ &= \mathbb{P}\Big(Y \in \bigcup_{\pi \in \mathcal{S}_n} \pi(B \cap O)\Big) = \sum_{\pi \in \mathcal{S}_n} \mathbb{P}\big(Y \in \pi(B \cap O)\big).\end{aligned}$$

Since exchangeability implies that for every permutation $\pi \in \mathcal{S}_n$ it holds that $\mathbb{P}(Y \in \pi(B \cap O)) = \mathbb{P}(Y \in B \cap O)$, we obtain that

$$\mathbb{P}\big(Y_{(\cdot),n} \in B\big) = n!\,\mathbb{P}\big(Y \in \pi(B \cap O)\big), \quad \text{for every } \pi \in \mathcal{S}_n. \tag{6}$$

Using the result in (6), we have

$$\begin{aligned}&\mathbb{P}\big((Y_{\pi(1)}, \ldots, Y_{\pi(n)}) \in O, Y_{(\cdot),n} \in B\big) \\ &\qquad = \mathbb{P}\big((Y_1, \ldots, Y_n) \in \pi^{-1}(O), Y_{(\cdot),n} \in B \cap O\big) \\ &\qquad = \mathbb{P}\big((Y_1, \ldots, Y_n) \in \pi^{-1}(B \cap O)\big) \\ &\qquad = \frac{1}{n!}\,\mathbb{P}\big(Y_{(\cdot),n} \in B\big) = \mathbb{E}\Big(\frac{1}{n!}\,\mathbb{1}_{\left\{ Y_{(\cdot),n} \in B \right\}}\Big).\end{aligned}$$

□

# 3 Conformal prediction: general theory

## 3.1 Main definitions

In this section, we introduce conformal prediction in the general case with all supplementary definitions and prove two important results: strong validity and optimality of conformal intervals.

In this subsection, we will follow Section 2 and Section 3 from [2].

Notice that we will work with the so-called "on-line protocol", in which we assume that the data is generated sequentially unlike more classical "batch" approach, where the number of data points is fixed. According to the "on-line protocol", we will also sequentially construct prediction intervals and explore the statistical relation between errors on each step.

From the mathematical perspective, we have an infinite sequence of random vectors $\big(X_n, Y_n\big)$, $n \geq 1$, which takes values in the measurable space $\big(\mathbb{R}^K \times \mathbb{R}\big)^\infty$ with $\sigma$-algebra $\mathcal{F}$. The true distribution $\mathbb{P}$ of $\big(X_n, Y_n\big)$, $n \geq 1$ belongs to a set of distributions $\mathcal{P}$. For convenience, we will denote random data points $(X_n, Y_n)$ as $Z_n$.

Recall that our task is for every $n \geq 2$, every sequence of "previous" data points $(X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1}) \in \mathbb{R}^K \times \mathbb{R}$, value $X_n$ for a new observation, and every significance rate $\varepsilon \in (0, 1)$ to construct a measurable set

$$\Gamma^\varepsilon_n = \Gamma^\varepsilon_n\big((X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1}), X_n\big) \subset \mathbb{R}$$

such that the probability of covering the target variable $Y_n$, i.e. the probability of the event $Y_n \in \Gamma^\varepsilon_n$, is equal to $1 - \varepsilon$. The latter property of prediction intervals is called (exact) validity.

Note that if, in addition, the events $\big\{\, Y_n \in \Gamma^\varepsilon_n \,\big\}, n \geq 2$, are independent, then we say that the sequence of confidence predictors $\Gamma^\varepsilon_n$, $n \geq 2$, is strongly valid.

To introduce the concept of conformal prediction interval, we need to introduce some additional notions first. In the following definitions, we assume that $S_n$, $n \geq 1$, is a sequence of measurable functions $S_n : \big(\mathbb{R}^K \times \mathbb{R}\big)^n \to \Sigma_n$, where $\big(\Sigma_n, \mathfrak{C}_n\big), n \geq 1$, are measurable spaces.

**Definition 3.1.** A sequence of statistics $S_n$, $n \geq 1$, is said to be algebraically transitive if there exists a sequence of measurable functions $F_n : \Sigma_{n-1} \times \big(\mathbb{R}^K \times \mathbb{R}\big) \to \Sigma_n$, $n \geq 2$, such that

$$S_n(z_1, \ldots, z_n) = F_n\big(S_{n-1}(z_1, \ldots, z_{n-1}), z_n\big)$$

for all $z_1, \ldots, z_n \in \big(\mathbb{R}^K \times \mathbb{R}\big)^n$. Informally, the function $F_n$ allows to find the summary for the first $n$ data points simply updating the summary of the first $n-1$ data points with a new element $z_n$.

**Definition 3.2.** A sequence of statistics $S_n, n \geq 1$, is said to be totally sufficient for a set of probability distributions $\mathcal{P}$ if for every $n \geq 1$

- $S_n$ is sufficient for the corresponding sequence of random variables $Z_1, \ldots, Z_n$

- $Z_1, \dots, Z_n$ and $Z_{n+1}, \dots$ are conditionally independent given $S_n$. Informally,

$$\mathcal{L}\big(Z_1, \dots, Z_n \mid S_n, Z_{n+1}, \dots\big) = \mathcal{L}\big(Z_1, \dots, Z_n \mid S_n\big).$$

*Remark.* A sequence of statistics $S_n, n \geq 2$, is called an ATTS-sequence if it is both algebraically transitive and totally sufficient.

**Definition 3.3.** A nonconformity measure $A_n = A_n\big(s_{n-1}, (x_n, y_n)\big)$ for $n \geq 2$ is any measurable function $A_n : \Sigma_{n-1} \times \big(\mathbb{R}^K \times \mathbb{R}\big) \to \mathbb{R}$.

A nonconformity measure is a function that computes how different the previous data (given through its summary) are from a new data point. Despite the generality of this definition, in our next examples large values of nonconformity measure will always mean large differences between the corresponding objects. However, note that it follows from Theorem 3.1 that conformal prediction produces valid prediction intervals for any choice (and interpretation) of nonconformity measure.

To construct conformal prediction intervals, we will apply a classical statistical technique of building confidence intervals for some random variable using its $p$-values (see Subsection 3.4). In our case, we will consider $p$-values of nonconformity measures $A_n$ conditionally on summary statistic $S_n$.

**Definition 3.4.** For the data points $z_1, \dots, z_n \in \big(\mathbb{R}^K \times \mathbb{R}\big)^n$ and the number $t_n \in [0, 1]$, define the (randomized) $p$-value as the function (for details see Subsection 3.4)

$$p_n\big(z_1, \dots, z_n, t_n\big) = \mathbb{P}\big(A_n > a_n \mid S_n = s_n\big) + t_n \mathbb{P}\big(A_n = a_n \mid S_n = s_n\big), \tag{7}$$

where $s_n = S_n\big(z_1, \dots, z_n\big)$ and $a_n = A_n\big(S_{n-1}(z_1, \dots, z_{n-1}), z_n\big)$.

*Remark.* On the right-hand side of the relation (7), we consider the distribution of the nonconformity measure $A_n = A_n(S_{n-1}, Z_n)$ on the space of pairs

$$F^{-1}(s_n) = \big\{(s_{n-1}, z_n) \in \Sigma_{n-1} \times \big(\mathcal{X} \times \mathcal{Y}\big) \mid F(s_{n-1}, z_n) = s_n\big\}.$$

In this thesis, we define a randomized $p$-value of a random variable $X$ as a function $p_X(x, t) = \mathbb{P}(X > x) + t\,\mathbb{P}(X = x)$, where $t \in [0, 1]$, i.e., it is a function that describes the distribution of a random variable $X$ similarly to the cumulative distribution function. Hence in case of the $p$-value in (7), it will later be convenient to consider it in a more classical way, namely as

$$p_n = p_{A_n|S_n}(a_n, t_n \mid s_n).$$

*Remark.* Since algebraic transitivity implies that $S_n = F_n(S_{n-1}(z_1, \dots, z_{n-1}), z_n)$, we can see that the $p$-value (7) depends on $z_1, \dots, z_n$ through the values of $S_{n-1}$ and $z_n$. Hence $p_n(z_1, \dots, z_n, t_n) = p_n(S_{n-1}, z_n, t_n)$.

Recall that we consider a hypothetical value $y \in \mathbb{R}$ of the target variable $Y_n$. Using $p$-value (7), we can compute how probable it is to get a larger nonconformity than the nonconformity between the summary $S_{n-1}$ of observation $z_1, \dots, z_{n-1}$ and the hypothetical next observation $z_n = (x_n, y)$. If this probability is too small, then it is natural not to include it in the prediction interval. This idea is realized in the definition of a conformal interval.

**Definition 3.5.** Assume that $Z_n$, $n \geq 1$ is a random sequence from $\big(\mathbb{R}^K \times \mathbb{R}\big)^\infty$ with some distribution $\mathbb{P}$ from the set $\mathcal{P}$. Let $S_n$, $n \geq 1$, be an ATTS sequence of statistics for the sequence $Z_n$, $n \geq 1$, and let $p_n$, $n \geq 2$, be the sequence of the corresponding $p$-values for nonconformity measures $A_n, n \geq 2$.

Suppose that the data points $z_1, \ldots, z_{n-1}$ and the explanatory variable $x_n$ of the next observation are known. Then for the number $t_n \in [0,1]$ and confidence level $\varepsilon \in [0,1]$, define a conformal prediction interval as

$$\Gamma_n^\varepsilon = \Gamma_n^\varepsilon\big(z_1, \ldots, z_{n-1}, x_n, t_n\big) = \big\{y \in \mathbb{R} \mid p_n(z_1, \ldots, z_{n-1}, (x_n, y), t_n) \geq \varepsilon\big\}.$$

*Remark.* Let $g$ be a strictly increasing function. If we fix a value $s_n$ of ATTS-statistic $S_n$, then for any two pairs $(s_{n-1}, z_n)$ and $(s'_{n-1}, z'_n)$ such that $s_n = F(s_{n-1}, z_n) = F(s'_{n-1}, z'_n)$, it holds that

$$A_n(s_{n-1}, z_n) > A_n(s'_{n-1}, z'_n) \iff g\big(A_n(s_{n-1}, z_n)\big) > g\big(A_n(s'_{n-1}, z'_n)\big).$$

Therefore, we get

$$\mathbb{P}\big(A_n > a_n \mid S_n = s_n\big) = \mathbb{P}\big(g(A_n) > g(a_n) \mid S_n = s_n\big).$$

It follows that conformal intervals constructed using the nonconformity measure $A_n$ are equal to the conformal intervals constructed using the measure $g(A_n)$.

## 3.2 Strong validity of conformal intervals

In this subsection, we consider in detail Theorem 1 in [2].

**Theorem 3.1** (Joint distribution of $p$-values)**.** In addition to the setting of Definition 3.5, assume that $\tau_n$, $n \geq 2$, is a sequence of independent and $U([0,1])$-distributed random variables, which is also independent of $Z_n$, $n \geq 2$.

Then the $p$-values

$$p_n\big(Z_1, \ldots, Z_n, \tau_n\big),\ n \geq 2$$

are independent and $U([0,1])$-distributed random variables.

*Proof.* 1) Distribution of $p$-values

According to Theorem 3.3 in Section 3.4, if $X$ is a random variable and $\tau$ is a $U([0,1])$-distributed random variable independent of $X$ , then the $p$-value $p_X(X, \tau)$ is $U([0,1])$-distributed.

Recall that $p_n = p_{A_n|S_n}(a_n, t_n \mid s_n)$ is a $p$-value of the random variable $A_n$ conditionally distributed given $S_n = s_n$. Hence, Theorem 3.3 implies that the conditional distribution of the random variable $p_{A_n|S_n}(A_n, \tau_n \mid S_n)$ given $S_n$ is uniform on $[0,1]$

Therefore, we have for every $\varepsilon \in [0,1]$ that $\mathbb{P}(p_n < \varepsilon \mid S_n) = \varepsilon$ a.s. Integrating over $S_n$ implies

$$\mathbb{P}\big(p_n < \varepsilon\big) = \mathbb{E}\,\mathbb{P}\big(p_n < \varepsilon \mid S_n\big) = \varepsilon. \tag{8}$$

2) Independence of $p$-values

For every $n \geq 2$ and sequence $\varepsilon_2, \ldots, \varepsilon_n \in [0,1]$, we want to prove that

$$\mathbb{E}\big(\mathbb{1}_{p_2<\varepsilon_2} \ldots \mathbb{1}_{p_n<\varepsilon_n}\big) = \varepsilon_2 \ldots \varepsilon_n.$$

Define a decreasing sequence of sub-$\sigma$-algebras by

$$\mathcal{G}_n = \sigma\big(S_n(Z_1, \ldots, Z_n), Z_{n+1}, \tau_{n+1}, \ldots\big), \text{ for } n \geq 2,$$

and consider first the conditional case w.r.t. $\mathcal{G}_n$, i.e., prove that

$$\mathbb{E}\big(\mathbb{1}_{p_2<\varepsilon_2} \ldots \mathbb{1}_{p_n<\varepsilon_n} \mid \mathcal{G}_n\big) = \varepsilon_2 \ldots \varepsilon_n, \text{ a.s.} \tag{9}$$

The latter statement we will prove by induction. For the case $n = 2$, we have due to the total sufficiency of $S_n$ and conditional distribution of $p_2$ given $S_2$, that

$$\mathbb{E}\big(\mathbb{1}_{p_2<\varepsilon_2} \mid \mathcal{G}_2\big) = \mathbb{E}\big(\mathbb{1}_{p_2<\varepsilon_2} \mid S_2\big) = \varepsilon_2.$$

Now suppose that the relation (9) holds for every $k < n$. Note that since $p_n$ depends on $Z_1, \ldots, Z_n$ through the values $S_{n-1}(Z_1, \ldots, Z_{n-1})$ and $Z_n$, we have that $p_n$ is $\mathcal{G}_{n-1}$-measurable. Hence, using the tower property, we get

$$\begin{aligned}
&\mathbb{E}\big(\mathbb{1}_{p_2<\varepsilon_2} \ldots \mathbb{1}_{p_n<\varepsilon_n} \mid \mathcal{G}_n\big) \\
&\qquad = \mathbb{E}\big(\mathbb{E}\big(\mathbb{1}_{p_2<\varepsilon_2} \ldots \mathbb{1}_{p_n<\varepsilon_n} \mid \mathcal{G}_{n-1}\big) \mid \mathcal{G}_n\big) \\
&\qquad = \mathbb{E}\big(\mathbb{1}_{p_n<\varepsilon_n} \mathbb{E}\big(\mathbb{1}_{p_2<\varepsilon_2} \ldots \mathbb{1}_{p_{n-1}<\varepsilon_{n-1}} \mid \mathcal{G}_{n-1}\big) \mid \mathcal{G}_n\big) \\
&\qquad = \mathbb{E}\big(\mathbb{1}_{p_n<\varepsilon_n} \varepsilon_2 \ldots \varepsilon_{n-1} \mid \mathcal{G}_n\big) \\
&\qquad = \varepsilon_2 \ldots \varepsilon_{n-1}\, \mathbb{E}\big(\mathbb{1}_{p_n<\varepsilon_n} \mid \mathcal{G}_n\big) = \varepsilon_2 \ldots \varepsilon_n.
\end{aligned}$$

Integrating this relation, we obtain (9). □

*Remark.* Theorem 3.1 implies that randomized conformal intervals are strongly valid, since for every $n \geq 2$, we have for the event that $Y_n$ is covered by $\Gamma_n^\varepsilon$ , that

$$A_n = \big\{\, Y_n \in \Gamma_n^\varepsilon(Z_1, \ldots, Z_{n-1}, X_n, \tau_n) \,\big\} = \big\{\, p_n(Z_1, \ldots, Z_{n-1}, (X_n, Y_n), \tau_n) > \varepsilon \,\big\}$$

and random variables $p_n$, $n \geq 2$, are independent and $U([0,1])$-distributed.

*Remark.* Note that in the case of deterministic conformal intervals, i.e., if $\tau_n = 1$ a.s., we get larger $p$-values and therefore larger conformal intervals $\Gamma_n^\varepsilon(1)$. It is easy to see that in this case, the so-called conservative validity

$$\mathbb{P}\big(Y_n \in \Gamma_n^\varepsilon(Z_1, \ldots, Z_{n-1}, X_n, 1)\big) \geq 1 - \varepsilon$$

holds.

## 3.3 Optimality

The result presented in this subsection (Theorem 2 in [2]) states that for any prediction set from a large class there is a conformal interval which is not worse than a given prediction set.

**Definition 3.6.** A prediction set $\Gamma^+$ is said to be at least as good as the prediction set $\Gamma$ if for every $\varepsilon \in (0,1)$ there holds $\Gamma^{+,\,\varepsilon} \subseteq \Gamma^{\varepsilon}$ almost surely.

Recall that in the case of the exchangeability model, the conformal interval was defined as

$$\Gamma_n^{\varepsilon}(y_1, \ldots, y_{n-1}) = \Big\{\, y \in \mathbb{R} \mid \frac{\#\{i = 1, \ldots, n \mid A^{(i)} > A^{(n)}\}}{n} \geq \varepsilon \,\Big\}$$

with the nonconformity measure $A_n = |y_n - \overline{y}_{n-1}|$ and the summary statistic equal to the vector of order statistic. It is easy to see that for any permutation $\pi \in \mathcal{S}_n$, the value $\#\{i = 1, \ldots, n \mid A^{(i)} > A^{(n)}\}$ will not change. Therefore, we get

$$\Gamma_n^{\varepsilon}(x_1, \ldots, x_{n-1}) = \Gamma_n^{\varepsilon}(x_{\pi(1)}, \ldots, x_{\pi(n-1)}).$$

The latter implies that we have equal conformal intervals for any two sequences of previous observations with equal vectors of order statistics. This intuition leads to the following definition.

**Definition 3.7.** If the prediction set $\Gamma_n = \Gamma_n\big(z_1, \ldots, z_{n-1}, x_n\big)$ depends on the previous observations $z_1, \ldots, z_{n-1}$ only through the value of their summary statistic $S_{n-1}$, i.e. $\Gamma_n = \Gamma_n(S_{n-1}, x_n)$, then $\Gamma_n$ is said to be invariant with respect to the statistic $S_{n-1}$.

**Definition 3.8.** The statistic $S$ with values in a measurable space $\Sigma$ is called boundedly complete (for statistical model $\mathcal{P}$) if for every bounded measurable function $f : \Sigma \to \mathbb{R}$, it holds that

$$\begin{aligned} &\mathbb{E} f(S) = 0 \text{ under every distribution } \mathbb{P} \in \mathcal{P} \\ &\qquad \Longrightarrow f(S) = 0, \mathbb{P} - \text{a.s. for every } \mathbb{P} \in \mathcal{P}. \end{aligned}$$

**Theorem 3.2.** Let $S_n, n \geq 2$, be an ATTS sequence of statistics, where every $S_n$ is boundedly complete. Assume that for every $n \geq 2$, the prediction interval $\Gamma_n^{\varepsilon}$ is exactly valid for every $\varepsilon \in [0,1]$ and invariant with respect to the corresponding statistic $S_{n-1}$.

Then, for every $n \geq 2$, there exists a deterministic conformal predictor $\Gamma_n^{\varepsilon,+}$, which is at least as good as $\Gamma_n^{\varepsilon}$.

*Proof.* First note that bounded completeness of the statistic $S_n$ implies exact conditional validity of $\Gamma_n$ given $S_n$:

$$\forall \mathbb{P} \in \mathcal{P} : \quad \mathbb{P}\big(y_n \notin \Gamma_n^{\varepsilon}(S_{n-1}, x_n) \mid S_n\big) = \varepsilon \quad \mathbb{P}\text{-a.s.}$$

since for measurable and bounded function $g(S_n) = \mathbb{P}\big(y_n \notin \Gamma_n^{\varepsilon}(S_{n-1}, x_n) \mid S_n\big)$, it holds that

$$\mathbb{E}\mathbb{P}\big(y_n \notin \Gamma_n^{\varepsilon}(S_{n-1}, x_n) \mid S_n\big) = \mathbb{P}\big(y_n \notin \Gamma_n^{\varepsilon}(S_{n-1}, x_n)\big) = \varepsilon$$

and finally $\mathbb{E}\Big(\mathbb{P}\big(y_n \notin \Gamma_n(S_{n-1}, x_n) \mid S_n\big) - \varepsilon\Big) = 0$.

To construct a conformal interval, we consider the following nonconformity measure defined using $\Gamma_n^\varepsilon$:

$$A_n\big(s_{n-1}, (x_n, y_n)\big) = 1 - \inf\big\{\, \varepsilon \in [0,1] \mid y_n \notin \Gamma_n^\varepsilon(s_{n-1}, x_n)\big\}. \tag{10}$$

Note that if the vector $(x_n, y_n)$ is very different from the previous observations, then to cover $(x_n, y_n)$ by $\Gamma_n^\varepsilon$, we need a wide interval, which corresponds to a small value of $\varepsilon$. As we can see in (10), a small value of $\varepsilon$ implies a large value of the nonconformity measure.

As we know, conformal intervals are constructed using the conditional distribution of the nonconformity measures given a summary statistic. In the proof of the theorem we will use the following facts.

*Claim 1.* For the nonconformity measure (10) and any $\delta \in (0,1)$, there holds the relation:

$$\mathbb{P}(A_n \geq 1 - \delta \mid S_n) \leq \delta.$$

*Proof.* (Claim 1) There holds

$$\begin{aligned}
\mathbb{P}(A_n \geq 1 - \delta \mid S_n) &= \mathbb{P}\big(\inf\big\{\, \varepsilon \in [0,1] \mid y_n \notin \Gamma_n^\varepsilon(s_{n-1}, x_n)\big\} \leq \delta \mid S_n\big) \\
&= \lim_{\gamma \downarrow \delta} \mathbb{P}\big(\inf\big\{\, \varepsilon \in [0,1] \mid y_n \notin \Gamma_n^\varepsilon(s_{n-1}, x_n)\big\} < \gamma \mid S_n\big) \\
&\leq \lim_{\gamma \downarrow \delta} \mathbb{P}\big(y_n \notin \Gamma_n^\gamma \mid S_n\big) = \lim_{\gamma \downarrow \delta} \gamma = \delta.
\end{aligned}$$

□

*Claim 2.* Assume that for the nonconformity measure (10) and for $\gamma, \delta \in (0,1)$, there holds $\mathbb{P}(A_n \geq \gamma \mid S_n) > \delta$. Then we have that $\gamma < 1 - \delta$.

*Proof.* (Claim 2)

If $\gamma \geq 1 - \delta$, then Claim 1 implies that

$$\mathbb{P}(A_n \geq \gamma \mid S_n) \leq \mathbb{P}(A_n \geq 1 - \delta \mid S_n) \leq \delta.$$

This gives a contradiction. □

Now define a conformal interval using the nonconformity measure (10),

$$\Gamma_n^{\varepsilon,+} = \big\{y \in \mathbb{R} \mid \mathbb{P}(A_n \geq A_n(s_{n-1}, (x_n, y)) \mid S_n) > \varepsilon\big\},$$

and consider the relation between our initial prediction interval $\Gamma_n^\varepsilon$ and its improvement, i.e., the conformal interval $\Gamma_n^{\varepsilon,+}$:

$$\begin{aligned}
y \in \Gamma_n^{\varepsilon,+} &\Longleftrightarrow \mathbb{P}\big(A_n \geq A_n(s_{n-1}, (x_n, y)) \mid S_n\big) > \varepsilon \\
(Claim\,2) &\Longrightarrow A_n(s_{n-1}, (x_n, y)) < 1 - \varepsilon \\
&\Longrightarrow \inf\big\{\, \gamma \in [0,1] \mid y_n \notin \Gamma_n^\gamma(s_{n-1}, x_n)\big\} > \varepsilon \\
&\Longrightarrow y \in \Gamma_n^\varepsilon.
\end{aligned}$$

Thus $\Gamma_n^{\varepsilon,+} \subset \Gamma_n^{\varepsilon}$.

To finish the proof, we need to show that for the constructed deterministic conformal interval $\Gamma_n^{\varepsilon,+}$, it holds that $\mathbb{P}\big(y_n \notin \Gamma_n^{\varepsilon,+}\big) = \varepsilon$. From the relation $\Gamma_n^{\varepsilon,+} \subset \Gamma_n^{\varepsilon}$ we get $\mathbb{P}\big(y_n \notin \Gamma_n^{\varepsilon,+}\big) \geq \varepsilon$. On the other hand, the remark to Theorem 3.1 implies that $\mathbb{P}\big(y_n \notin \Gamma_n^{\varepsilon,+}\big) \leq \varepsilon$.

□

## 3.4 p-values

In statistics, a $p$-value is a quantity which measures the probability of obtaining a more extreme result produced by a random variable than what we observe. In this section, we define a $p$-value as a function of the observed value and prove an important result that explains the procedure of constructing confidence intervals using $p$-values.

**Definition 3.9.** Let $X$ be a random variable. Define the deterministic $p$-value to be a function

$$p_X(x) = \mathbb{P}(X \geq x),\ x \in \mathbb{R}$$

and the randomized $p$-value as

$$p_X(x,\, t) = \mathbb{P}(X > x) + t\mathbb{P}(X = x),\ x \in \mathbb{R},\ t \in [0, 1].$$

*Remark.* Despite the function $p_X(x, t)$ defined above being deterministic, we call it randomized since in this and the next sections we will mostly discuss results for the random function $p_X(x, \tau)$, where $\tau$ is a $U([0, 1])$-distributed random variable.

For a random variable $X$ with deterministic and strictly decreasing $p$-value natural ways of constructing a confidence interval with a confidence level equal to $1 - \varepsilon \in (0, 1)$ are, for example, to "cut off" a tail with a mass $\varepsilon$ from one side or two tails of mass $\varepsilon/2$ each from both sides:

$$C_1^{\varepsilon} = \big\{\, x \in \mathbb{R} \mid p_X(x) > \varepsilon \,\big\} \subset \mathbb{R},$$

$$C_2^{\varepsilon} = \big\{\, x \in \mathbb{R} \mid p_X(x) \in [\varepsilon/2, 1 - \varepsilon/2] \,\big\} \subset \mathbb{R}.$$

The following result allows us to construct confidence intervals also for the case of general $p$-values, in which non-continuous or non-increasing cases are included.

**Theorem 3.3.** Let $X$ and $\tau$ be independent random variables, $\tau \sim U([0, 1])$. Then for the randomized $p$-value of $X$ it holds that

$$p_X(X, \tau) \sim U([0, 1]). \tag{11}$$

*Proof.* Define a randomized cumulative distribution function of $X$ by

$$F_X(x, t_0) = \mathbb{P}(X < x) + t_0\mathbb{P}(X = x),\ \ x \in \mathbb{R}, t_0 \in [0, 1].$$

Even though the function $F_X(x, t_0)$ is deterministic, we call it randomized, since we will use a $U([0, 1])$-distributed random variable $\tau$ instead of the parameter $t_0$.

To simplify the notation in the proof of the theorem, we will use $F(x)$ and $F(x,t)$ instead of $F_X(x)$ and $F_X(x,t)$ respectively.

Note that if $t_0 = 1 - t$, then $p_X(x,t) = 1 - F(x,t_0)$. It means it is enough to prove that

$$F(X,\tau_0) \sim U([0,1]) \tag{12}$$

for $\tau_0 \sim U([0,1])$, i.e., for every $\varepsilon \in [0,1]$ we need to show that $\mathbb{P}\big(F(X,\tau_0) \leq \varepsilon\big) = \varepsilon$.

Before we move to the proof of (12), we state two useful results about deterministic and randomized cumulative distribution functions.

*Claim* 1. For every $x \in \mathbb{R}$ and $t_0 \in [0,1]$, it holds that $F(x,t_0) \leq F(x)$.

*Claim* 2. Assume that for $\varepsilon \in (0,1)$, the preimage $F^{-1}(\{\varepsilon\})$ is non-empty. Then for every $x > \sup F^{-1}(\{\varepsilon\})$, it holds that $\mathbb{P}(X < x) > \varepsilon$.

*Proof.* (Claim 2)

For any $x' \in \big(\sup F^{-1}(\{\varepsilon\}),\, x\big)$, we have that $\varepsilon < F(x') \leq \mathbb{P}(X < x)$. □

Now we turn to the proof of the main theorem.

**Case 1**: $F^{-1}(\{\varepsilon\}) = \{x_\varepsilon\}$

Claim 2 implies that for every $x > x_\varepsilon$, we have $\mathbb{P}(X < x) > \varepsilon$ and therefore $F(x,t_0) > \varepsilon$ for every $t_0 \in [0,1]$. Hence

$$\big\{\, F(X,\tau_0) \leq \varepsilon \,\big\} \subset \big\{\, X \leq x_\varepsilon \,\big\}.$$

From Claim 1, we get that for every $x \leq x_\varepsilon$ and $t_0 \in [0,1]$, there holds

$$F(x,t_0) \leq F(x) \leq F(x_\varepsilon) = \varepsilon,$$

which implies that $\big\{\, X \leq x_\varepsilon \,\big\} \subset \big\{\, F(X,\tau_0) \leq \varepsilon \,\big\}$.

Hence, we have that $\big\{\, X \leq x_\varepsilon \,\big\} = \big\{\, F(X,\tau_0) \leq \varepsilon \,\big\}$ and finally that

$$\mathbb{P}(F(X,\tau_0) \leq \varepsilon) = \mathbb{P}(X \leq x_\varepsilon) = \varepsilon.$$

**Case 2**: $\#\, F^{-1}(\{\varepsilon\}) = \infty$.

Here we consider 2 cases.

1) If $\mathbb{P}(X = \sup F^{-1}(\{\varepsilon\})) = 0$, we have $\big\{F(X,\tau_0) \leq \varepsilon\big\} = \big\{X \leq \sup F^{-1}(\{\varepsilon\})\big\}$ and therefore

$$\mathbb{P}(F(X,\tau_0) \leq \varepsilon) = \mathbb{P}(X \leq \sup F^{-1}(\{\varepsilon\})) = \varepsilon.$$

2) If $\mathbb{P}(X = \sup F^{-1}(\{\varepsilon\})) > 0$, we have

$$\big\{\, F(X,\tau_0) \leq \varepsilon \,\big\} = \big\{\, X < \sup F^{-1}(\{\varepsilon\}) \,\big\} \cup \big\{\, X = \sup F^{-1}(\{\varepsilon\}), \tau_0 = 0 \,\big\}.$$

Since $\mathbb{P}(X = \sup F^{-1}(\{\varepsilon\}), \tau_0 = 0) = 0$ and $\mathbb{P}(X < \sup F^{-1}(\{\varepsilon\})) = \varepsilon$, we obtain

$$\mathbb{P}(F(X,\tau_0) \leq \varepsilon) = \mathbb{P}(X < \sup F^{-1}(\{\varepsilon\})) + \mathbb{P}(X = \sup F^{-1}(\{\varepsilon\}), \tau_0 = 0) = \varepsilon.$$

**Case 3:** $F^{-1}(\{\varepsilon\}) = \emptyset$

Let $x_\varepsilon = \inf\big\{\, x \mid F(x) > \varepsilon \,\big\}$.

Since for $x > x_\varepsilon$ it holds that $\mathbb{P}(X < x) \geq \mathbb{P}(X \leq x_\varepsilon) > \varepsilon$, we get

$$\big\{\, F(X, \tau_0) \leq \varepsilon \,\big\} \subset \big\{\, X \leq x_\varepsilon \,\big\}.$$

More precisely,

$$\big\{\, F(X, \tau_0) \leq \varepsilon \,\big\} = \big\{\, X < x_\varepsilon \,\big\} \cup \big\{\, X = x_\varepsilon, \tau_0 \in \big[\, 0, (\varepsilon - \mathbb{P}(X < x_\varepsilon))/\mathbb{P}(X = x_\varepsilon) \,\big] \,\big\}.$$

Now using the independence of $X$ and $\tau_0$ together with the the fact that $\tau_0$ is uniformly distributed on $[0, 1]$, we get

$$\begin{aligned}
&\mathbb{P}\big(F(X, \tau_0) \leq \varepsilon\big) \\
&\qquad = \mathbb{P}\big(X < x_\varepsilon\big) + \mathbb{P}\big(X = x_\varepsilon, \tau_0 \in \big[\, 0, (\varepsilon - \mathbb{P}(X < x_\varepsilon))/\mathbb{P}(X = x_\varepsilon) \,\big]\big) \\
&\qquad = \mathbb{P}\big(X < x_\varepsilon\big) + \mathbb{P}\big(X = x_\varepsilon\big)\mathbb{P}\big(\tau_0 \in \big[\, 0, (\varepsilon - \mathbb{P}(X < x_\varepsilon))/\mathbb{P}(X = x_\varepsilon) \,\big]\big) \\
&\qquad = \mathbb{P}\big(X < x_\varepsilon\big) + \mathbb{P}\big(X = x_\varepsilon\big)\frac{\varepsilon - \mathbb{P}(X < x_\varepsilon)}{\mathbb{P}(X = x_\varepsilon)} = \varepsilon.
\end{aligned}$$

□

As a direct consequence of the previous theorem, we obtain the following statement.

**Corollary 3.3.1.** Let $X$ be a random variables and let $p(x, t)$, $x \in \mathbb{R}$, $t \in [0, 1]$ be a randomized $p$-value for $X$. If $\tau$ is $U([0, 1])$-distributed and independent of $X$, then for every $\varepsilon \in [0, 1]$ and every interval $(a, b) \subset [0, 1]$ of length $1 - \varepsilon$, the randomized confidence interval $C^\varepsilon$, given by

$$C^\varepsilon = C^\varepsilon(\tau) = \big\{x \in \mathbb{R} \mid p_X(x, \tau) \in (a, b)\big\},$$

has significance level $\varepsilon$, i.e., $\mathbb{P}(X \in C^\varepsilon(\tau)) = 1 - \varepsilon$.

# 4 Conformal prediction for regression models

## 4.1 IID model

In this section, we consider results presented in Section 2.3 of [1] and in Section 5 of [2]. We provide the missing theory on bags in Appendix 4.1.1.

In this overview of applications of conformal prediction to regression models, we start with the IID model, where the only assumption is that the random variables $(x_n, y_n), n \geq 1$, are independent and identically distributed according to some unknown distribution.

For each model discussed in Section 4, we follow the same scheme: First we define an ATTS sequence of summary statistics for the considered model and then we use a nonconformity measure, whose conditional distribution given the summary statistic can be computed analytically.

As we can see from Theorem 4.3 in Appendix 4.1.1, bags of observations

$$S_n(z_1, \ldots, z_n) = \wr z_1, \ldots, z_n \int, n \geq 1,$$

form an ATTS sequence of statistics in the IID model. Basically, a sequence of bags is ATTS also for the exchangeability model. Therefore, all results obtained in this section hold also for the exchangeability model.

One of the natural ways to define a nonconformity measure is to compute the distance from a hypothetical value $y \in \mathbb{R}$ to a point prediction $\hat{y}_n$ obtained by applying any algorithm (linear regression, neural network, etc.) to the previous data:

$$A_n\big(s_{n-1}, (x_n, y_n)\big) = |y_n - \hat{y}_n|.$$

For the sake of simplicity, we consider the distance to the prediction given by ridge regression with parameter $a > 0$ (even though we do not have linearity assumption):

$$A_n\big(s_{n-1}, (x_n, y_n)\big) = |e_n| = \big|y_n - x_n'\big(X_n'X_n + aI_n\big)^{-1}X_n'Y_n\big|,$$

where $X_n$ is a matrix with rows $x_1', \ldots, x_n'$ and $Y = [\, y_1, \ldots, y_n \,]'$.

*Remark.* The ridge coefficient $a > 0$ ensures that the matrix $X_n'X_n + aI_n$ is not singular, since for any vector $x$ we have

$$\big(X_n'X_n + aI_n\big)x = 0 \;\Rightarrow\; x'\big(X_n'X_n + aI_n\big)x = 0 \;\Rightarrow\; ||X_nx||^2 + a||x||^2 = 0.$$

Since $a > 0$, it follows that $x = 0$.

Now consider the conditional distribution of the nonconformity measure $A_n$, given summary statistic $S_n$. According to Theorem 4.3, for exchangeable (in particular i.i.d.) random variables $z_1, \ldots, z_n$ the conditional distribution given the bag $\wr z_1, \ldots, z_n \int$ is uniform over all permutations of the data in the bag. Together with the fact that the value of the nonconformity measure $A_n$ is invariant with respect to the permutations of the previous observations, this implies that, conditionally on $S_n$ and almost surely, the nonconformity measure $A_n$ is uniformly distributed over

the items in the bag $\wr|e_1|, \ldots, |e_n|\int$, in the sense that each of $n$ items in the bag is assigned a probability $\frac{1}{n}$.

Recall that in conformal prediction, we test the nonconformity for every hypothetical value $y \in \mathbb{R}$ of the response variable $y_n$. Let $y_n = y$ and consider the vector of residuals

$$\begin{aligned} E_n &= Y_n - X_n\big(X_n'X_n + aI_n\big)^{-1}X_n'Y_n \\ &= U\,[\,y_1, \ldots, y_{n-1}, y\,]' \\ &= U\,[y_1, \ldots, y_{n-1}, 0]' + U\,[0, \ldots, 0, y]' \\ &= A + By, \end{aligned}$$

where $U = I_n - X_n\big(X_n'X_n + aI_n\big)^{-1}X_n'$ and $A, B$ are matrices. It follows that every residual $e_i$ depends linearly on $y$, i.e., that $e_i = e_i(y) = a_i + b_i y$ for every $i = 1, \ldots, n$, where the coefficients $a_i, b_i$ are fixed given a bag of $n$ observations.

Finally, using Theorem 4.3, we obtain the following expression for (deterministic) conformal intervals in the IID model:

$$\begin{aligned} \Gamma_n^\varepsilon &= \{\, y \in \mathbb{R} \mid p_n(y) > \varepsilon \,\} \\ &= \Big\{ y \in \mathbb{R} \mid \frac{1}{n}\#\{i = 1, \ldots, n \mid |e_i(y)| \geq |e_n(y)|\} > \varepsilon \Big\} \\ &= \Big\{ y \in \mathbb{R} \mid \frac{1}{n}\#\{i = 1, \ldots, n \mid |a_i + b_i y| \geq |a_n + b_n y|\} > \varepsilon \Big\}. \end{aligned}$$

This representation of the conformal interval provides us with an efficient algorithm for computing it. Note that function $|a_i + b_i y|$ can define either a union of two non-horizontal rays ($b_i \neq 0$) or a horizontal line ($b_i = 0$). Intersection points of the corresponding lines define a partition of the real line into a disjoint union of (finite and infinite) intervals. On every such interval, the $p$-value remains unchanged.

Hence, to compute the conformal interval, we need

1. to find solutions of the equations $|a_i + b_i y| = |a_n + b_n y|, i = 1, \ldots, n-1$,
2. for every interval defined by two neighboring solutions, to count the $p$-value
$$\#\{i = 1, \ldots, n \mid |a_i + b_i y| \geq |a_n + b_n y|\} \quad \text{and} \tag{13}$$
3. if the $p$-value (13) is larger than $\varepsilon$, then include the interval in the conformal interval.

The resulting interval is conservative in the sense that $\mathbb{P}(y_n \in \Gamma_n^\varepsilon) \geq \varepsilon$. Using randomized $p$-values, exact coverage can be obtained.

*Remark* (on computational aspects of conformal prediction). One of the disadvantages of conformal prediction is that we need to test every hypothetical value $y \in \mathbb{R}$, and exact analytical computation of conformal intervals is often not possible. Therefore, we have to build an approximation by considering the grid of the real line.

Note that to define a nonconformity measure of the form $A_n = |y_n - \hat{y}_n|$, we may use any method which provides a point predictor $\hat{y}_n$. Using the approach described in this section, we will obtain (conservatively) valid conformal intervals.

### 4.1.1 Appendix: Bags

In this appendix, we present the concept of a bag, which plays the role of a summary statistic for the IID model.

**Definition 4.1.** A *bag* $\wr x_1, \ldots, x_n \int$ is a collection of $n$ elements $x_1, \ldots, x_n$ such that

1. some of the elements may be equal.
2. permutations of elements do not change the bag.

Let $\mathcal{X}$ be a set. On the set $\mathcal{X}^n$ we consider the following relation: For $x, y \in \mathcal{X}^n$, let $x \sim y$ if there exists a permutation $\pi \in \mathcal{S}_n$ such that $x = \pi(y)$, where $\pi(y) = (y_{\pi(1)}, \ldots, y_{\pi(n)})$. By general properties of permutations, this is an equivalence relation. Hence, the set $\mathcal{X}^n$ can be represented as a disjoint union of the equivalence classes $\big\{ \big[ x \big], x \in \mathcal{X}^n \big\} = \mathcal{X}^n|_\sim$. For a subset $A \subset \mathcal{X}^n$, we denote $\big[ A \big] = \big\{ \big[ x \big] \mid x \in A \big\}$.

It is clear that between bags consisting of $n$ elements from $\mathcal{X}$ and the equivalence classes mentioned above there exists a bijective mapping: $\wr x_1, \ldots, x_n \int \longleftrightarrow \big[ x \big]$ where $x = (x_1, \ldots, x_n)$. Therefore, if we define a $\sigma$-algebra on the set of equivalence classes, we automatically obtain a $\sigma$-algebra on the set of bags.

**Definition 4.2.** Consider a set $\mathcal{X}$. For a subset $A \subset \mathcal{X}^n$, a *permutational closure* $\mathrm{pcl}(A) \subset \mathcal{X}^n$ is defined as

$$\mathrm{pcl}(A) = \big\{ \pi(x) \mid x \in A, \pi \in \mathcal{S}_n \big\}.$$

**Lemma 4.1.** *(properties of permutational closure)*

1. $A \subset pcl(A)$.
2. $\big[ A \big] = \big[ pcl(A) \big]$.

   *Let $\mathcal{F}$ be a $\sigma$-algebra on the set $\mathcal{X}$ and let $\mathcal{F}^n$ be the corresponding product $\sigma$-algebra on $\mathcal{X}^n$.*

3. *For $A \in \mathcal{F}^n$, we have that $pcl(A) \in \mathcal{F}^n$.*

*Proof.* 2) It holds that

$$\big[ x \big] \in \big[ A \big] \iff \exists y \in A \text{ s.t. } x \sim y \iff \exists \pi \in \mathcal{S}_n \text{ s.t. } x = \pi(y) \iff x \in \mathrm{pcl}(A).$$

3) Recall that a permutation of coordinates in $x \in \mathcal{X}^n$ is an $(\mathcal{F}^n, \mathcal{F}^n)$-measurable mapping. Now, since $\mathrm{pcl}(A) = \cup_{\pi \in \mathcal{S}_n} \pi(A)$ and $\pi(A) \in \mathcal{F}^n$ for every $\pi \in \mathcal{S}_n$, we have that $\mathrm{pcl}(A) \in \mathcal{F}^n$.

□

**Theorem 4.2.** Let $\mathcal{F}$ be a $\sigma$-algebra on a set $\mathcal{X}$ and let $\mathcal{F}^n$ be the corresponding product $\sigma$-algebra on the set $\mathcal{X}^n$. Then the class of subsets

$$\mathcal{C}^n = \big\{ \big[\, A \,\big] \mid A \in \mathcal{F}^n \big\}$$

is a $\sigma$-algebra on the set $\mathcal{X}^n|_\sim$.

*Proof.* 1) For any sequence $A_i \in \mathcal{F}^n$, $i \geq 1$, we have that

$$\bigcup_{i\geq 1} \big[\, A_i \,\big] = \big[\, \bigcup_{i\geq 1} A_i \,\big] \in \mathcal{C}^n,$$

since $\big[\, x \,\big] \in \cup_{i\geq 1} \big[\, A_i \,\big] \iff \exists\, i \geq 1 \,\exists\, y \in A_i$ s.t. $x \sim y \iff \exists\, y \in \cup_{i\geq 1} A_i$ s.t. $x \sim y \iff \big[\, x \,\big] \in \big[\, \cup_{i\geq 1} A_i \,\big]$.

2) For any $A \in \mathcal{F}^n$, we have

$$\big[\, A \,\big]^C = \big[\, \mathrm{pcl}(A)^C \,\big] \in \mathcal{C}^n,$$

since $\big[\, x \,\big] \notin \big[\, A \,\big] \iff$ for every $y \in A$ we have that $x \not\sim y \iff$ for every $y \in \mathrm{pcl}(A)$ there holds $x \not\sim y \iff \exists\, z \in \mathrm{pcl}(A)^C$ s.t. $x \sim z \iff \big[\, x \,\big] \in \big[\, \mathrm{pcl}(A)^C \,\big]$.

□

Now consider the mapping $g(x) = \big[\, x \,\big]$, $x \in \mathcal{X}^n$. It is easy to see that $g$ is $\big(\mathcal{F}^n, \mathcal{C}^n\big)$-measurable, since for every $\big[\, A \,\big] \in \mathcal{C}^n$, we have that

$$g^{-1}\big(\big[\, A \,\big]\big) = \mathrm{pcl}(A) \in \mathcal{F}^n.$$

*Remark.* As a consequence of the latter result, we also obtain that the class of permutational closures $\mathcal{G} = \big\{\, \mathrm{pcl}(A) \mid A \in \mathcal{F}^n \,\big\}$ is a sub-$\sigma$-algebra of $\mathcal{F}^n$.

For the next result, we need the notion of a uniform distribution on a bag. For $x \in \mathcal{X}^n$, the uniform distribution on $\wr x_1, \ldots, x_n \int$ is a discrete distribution which gives probability $\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{x_i = z\}$ to each value $z \in \{x_1, \ldots, x_n\}$. We will denote this distribution by $U(\wr x \int)$ or $U(\wr x_1, \ldots, x_n \int)$. For example, $U(\wr (1,2,3) \int)$ assigns probability $\frac{1}{3}$ to $1, 2$ and $3$ each; $U(\wr (1,1,2) \int)$ assigns probability $\frac{2}{3}$ and $\frac{1}{3}$ to 1 and 2, respectively; and $U(\wr (1,1,1) \int)$ is point-mass 1 at 1.

We will also consider all possible orderings of the components $x_i$, i.e., $(\pi(x))_{\pi\in\mathcal{S}_n} \in (\mathcal{X}^n)^{\mathcal{S}_n}$, and the uniform distribution on the corresponding bag, i.e., $U(\wr \pi(x), \pi \in \mathcal{S}_n \int)$, which puts probability $\frac{1}{n!} \sum_{\pi\in\mathcal{S}_n} \mathbb{1}\{\pi(x) = z\}$ to each value $z \in \{\pi(x) : \pi \in \mathcal{S}_n\}$. For example, $U(\wr \pi(1,2,3) : \pi \in \mathcal{S}_3 \int)$ puts probability $\frac{1}{6}$to each of the values $(1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), (3,2,1)$; $U(\wr \pi(1,1,2) : \pi \in \mathcal{S}_3 \int)$ puts probability $\frac{1}{3}$ to each of the values $(1,1,2), (1,2,1)$ and $(2,1,1)$; finally, $U(\wr \pi(1,1,1) : \pi \in \mathcal{S}_3 \int)$ is point-mass 1 at $(1,1,1)$.

**Theorem 4.3.** Consider exchangeable random variables $Z_1, \ldots, Z_n$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ taking values in $\mathcal{X}$. For $Z = (Z_1, \ldots, Z_n)$, the conditional distribution of $Z$ given $\wr Z \int$ is $U(\wr \pi(Z) : \pi \in \mathcal{S}_n \int)$. In particular, for an event of the form $\{Z \in A\}$ and almost all $\omega \in \Omega$, we have

$$\mathbb{P}\big(Z \in A \mid \wr Z \int\big)(\omega) = \frac{1}{n!} \sum_{\pi\in\mathcal{S}_n} \mathbb{1}\{\pi(Z(\omega)) \in A\}$$

*Proof.* Consider any event of the form $\{Z \in A\}$ and any event of the form $\{\lbag Z \rbag \in B\}$. In view of the definition of conditional probabilities, we need to show that $\mathbb{P}(Z \in A, \lbag Z \rbag \in B)$ equals

$$\begin{aligned}\int_{\{\lbag Z \rbag \in B\}} \frac{1}{n!} \sum_{\pi \in \mathcal{S}_n} \mathbb{1}\{\pi(Z) \in A\} d\mathbb{P} &= \frac{1}{n!} \sum_{\pi \in \mathcal{S}_n} \mathbb{P}(\pi(Z) \in A, \lbag Z \rbag \in B) \\ &= \frac{1}{n!} \sum_{\pi \in \mathcal{S}_n} \mathbb{P}(\pi(Z) \in A, \lbag \pi(Z) \rbag \in B) \\ &= \mathbb{P}(Z \in A, \lbag Z \rbag \in B),\end{aligned}$$

where the second equality holds because $\lbag Z \rbag = \lbag \pi(Z) \rbag$ and the third equality holds because $\mathbb{P}(\pi(Z) \in A, \lbag \pi(Z) \rbag \in B) = \mathbb{P}(Z \in A, \lbag Z \rbag \in B)$ by exchangeability.

□

## 4.2 Gauss linear model

This section corresponds to Section 6 in [2] and Section 8.5 in [1]. Besides the discussion of the results, we provide rigorous proofs of some omitted details of Proposition 8.4 from [1].

The Gauss linear model is a classical textbook model: For the random variables $(x_n, y_n) \in \mathbb{R}^K \times \mathbb{R}$, $n \geq 1$, we assume that there holds the linear relation

$$y_n = x_n' \beta + \varepsilon_n \tag{14}$$

with the following additional assumptions:

1. The row vectors $x_n'$ are deterministic for every $n \geq 1$.
2. $\varepsilon_n, n \geq 1$, are independent random variables.
3. $\varepsilon_n \sim N(0, \sigma^2)$ for every $n \geq 1$.

Parameters $\beta$ and $\sigma^2$ are assumed to be unknown constants.

In this section, we will use the following notation:

$$Y_n = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X_n = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}, \quad \mathcal{E}_n = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

To avoid singular matrices we also assume that

4. matrix $X_n$ has full rank.

A standard result for the Gauss linear model that allows the construction of prediction intervals for the random variable $y_n$ given previous data points and the value of the explanatory variable $x_n$, states (see Theorem 4.7) that the statistic

$$T_n = \frac{y_n - \hat{y}_n^{(n-1)}}{\hat{\sigma}_{n-1} \sqrt{1 + x_n' (X_{n-1}' X_{n-1})^{-1} x_n}} \tag{15}$$

is $t_{n-K-1}$-distributed. In the latter result, $\hat{y}_n^{(n-1)}$ denotes the prediction of $y_n$ given previous data points $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$ and $x_n$, i.e.

$$\hat{y}_n^{(n-1)} = x_n'(X_{n-1}'X_{n-1})^{-1}X_{n-1}'Y_{n-1}.$$

Result (15) implies that, for every $\varepsilon \in (0,1)$, the interval

$$\mathcal{C}^\varepsilon = \left[\hat{y}_n^{(n-1)} \pm t_{n-K-1}^{1-\varepsilon/2}\,\hat{\sigma}_{n-1}\sqrt{1 + x_n'(X_{n-1}'X_{n-1})^{-1}x_n}\right] \tag{16}$$

contains $y_n$ with probability $1-\varepsilon$. The $\mathcal{C}^\varepsilon$ are called Student's prediction intervals.

In the rest of the section, we want to show that Student's prediction interval can be obtained by applying conformal prediction. As usual, we start with defining an ATTS sequence of statistics for the considered model.

**Lemma 4.4.** *The sequence of statistics*

$$S_n(X_n, Y_n) = \big(X_n, X_n'Y_n, Y_n'Y_n\big), n \geq 1, \tag{17}$$

*is ATTS for the Gauss linear model.*

*Proof.* 1) Algebraic transitivity.

We have that

$$X_n'Y_n = \sum_{i=1}^n y_i x_i = X_{n-1}'Y_{n-1} + y_n x_n \;\text{ and }\; Y_n'Y_n = Y_{n-1}'Y_{n-1} + y_n^2.$$

Hence, there exists a sequence of measurable functions $F_n$, $n \geq 2$, such that $S_n = F_n\big(S_{n-1}, (x_n, y_n)\big)$, for every $n \geq 2$.

2) Total sufficiency.

Here we will use a classical result about sufficient statistics called the Neyman-Fisher factorization theorem.

**Theorem 4.5.** (Neyman-Fisher factorization theorem)

Let $(P_\theta, \Theta)$ be a parametric set of absolutely continuous distributions for the random vector $X$. If the likelihood $L_\theta(x)$ can be represented as $h(x)g(T(x), \theta)$, where $h$ and $g$ are measurable functions, then $T(x)$ is a sufficient statistic with respect to the statistical model $(P_\theta,\ \Theta)$.

Now, since the random variables $y_1, \ldots, y_n$ are independent and $N(x_i'\beta, \sigma^2)$-distributed, we have:

$$\begin{aligned} L_{\beta,\sigma^2}(y_1, \ldots, y_n) &= \prod_{i=1}^n L_{\beta,\sigma^2}(y_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}\big(y_i - x_i'\beta\big)^2\Big) \\ &= \big(2\pi\sigma^2\big)^{-\frac{n}{2}} \exp\Big(-\frac{(Y_n - X_n\beta)'(Y_n - X_n\beta)}{2\sigma^2}\Big) \\ &= \big(2\pi\sigma^2\big)^{-\frac{n}{2}} \exp\Big(-\frac{Y_n'Y_n - 2\beta'X_n'Y_n + \beta'X_n'X_n\beta}{2\sigma^2}\Big). \end{aligned}$$

From the Neyman-Fisher factorization theorem, it follows that the sequence (17) is sufficient for Gauss linear model.

As discussed in the previous section, conditional independence follows from the assumption of independence. □

**Theorem 4.6.** Consider the Gauss linear model (14) together with the ATTS sequence of statistics $S_n = (X_n, X_n'Y_n, Y_n'Y_n)$, $n \geq 1$. The conformal prediction intervals constructed using nonconformity measures

$$A_n\big(S_{n-1}, (x_n, y_n)\big) = \frac{|y_n - \hat{y}_n^{(n-1)}|}{\hat{\sigma}_{n-1}\sqrt{1 + x_n'(X_{n-1}'X_{n-1})^{-1}x_n}} \quad \text{or}$$

$$A_n'\big(S_{n-1}, (x_n, y_n)\big) = |y_n - \hat{y}_n^{(n-1)}|$$

are equal to the Student's prediction interval (16).

*Proof.* We know from Theorem 4.7 that the nonconformity measure $A_n$ is unconditionally $|t_{n-K-1}|$-distributed. At the same time, it follows from Theorem 4.8 that $A_n$ and the summary statistic $S_n$ are independent. Therefore, $A_n$ is $|t_{n-K-1}|$-distributed also conditionally given $S_n$.

The latter fact implies that for a conformal interval $\Gamma_n^\varepsilon$ that is constructed using the nonconformity measure $A_n$, it holds that

$$\begin{aligned}
y \in \Gamma_n^\varepsilon(S_{n-1}, x_n) &\iff p_n\big(S_{n-1}, (x_n, y)\big) \geq \varepsilon \\
&\iff \mathbb{P}\big(A_n \geq A_n(S_{n-1}, (x_n, y)) \mid S_n = F_n(S_{n-1}, (x_n, y))\big) \geq \varepsilon \\
&\iff \mathbb{P}\big(A_n \geq A_n(S_{n-1}, (x_n, y))\big) \geq \varepsilon \\
&\iff A_n\big(S_{n-1}, (x_n, y)\big) \leq t_{n-K-1}^{1-\varepsilon/2}.
\end{aligned}$$

The last inequality defines the Student's prediction interval.

As mentioned in a remark in Subsection 3.1, to prove that two nonconformity measures $A_n$ and $A_n'$ produce equal conformal intervals, it is enough to show that $A_n' = g(A_n)$, where $g$ is a strictly increasing function when the value of $S_n$ is fixed.

From Theorem 4.13, we have that

$$y_n - \hat{y}_n^{(n-1)} = \frac{e_n}{1 - h_{nn}} \quad \text{and} \quad \hat{\sigma}_{n-1} = \sqrt{\frac{1}{n-K-1}\Big(\sum_{i=1}^{n} e_i^2 - \frac{e_n^2}{1-h_{nn}}\Big)},$$

which implies that

$$\frac{|y_n - \hat{y}_n^{(n-1)}|}{\hat{\sigma}_{n-1}\sqrt{1 + x_n'(X_{n-1}'X_{n-1})^{-1}x_n}} = \frac{|y_n - \hat{y}_n^{(n-1)}|}{B\sqrt{C - c(y_n - \hat{y}_n^{(n-1)})^2}}$$

where the values $B = \sqrt{\frac{1}{n-K-1}\big(1 + x_n'X_{n-1}'X_{n-1})^{-1}x_n\big)}$ and $C = \sum_{i=1}^{n} e_i^2$ are constant, since we assume that $S_n = (X_n, X_n'Y_n, Y_n'Y_n)$ is fixed.

Now consider the function $g(x) = \dfrac{x}{B\sqrt{C - cx^2}}$, $x > 0$, and its derivative

$$g'(x) = \frac{1}{B}\frac{\sqrt{C - cx^2} - x\frac{-2xc}{2\sqrt{C-cx^2}}}{C - cx^2} = \frac{1}{B}\frac{2(C - cx^2) + 2cx^2}{2(C - cx^2)\sqrt{C - cx^2}} = \frac{1}{B}\frac{C}{(C - cx^2)\sqrt{C - cx^2}}.$$

As we can see, the derivative $g'(x)$ is positive for all $x \in D(g)$. Hence, the function $g$ is strictly increasing, and conformal intervals produced by the nonconformity measure $A'_n$ are equal to the Student's prediction intervals. □

### 4.2.1 Appendix

**Distribution of T-statistic**

**Theorem 4.7.** In the Gauss linear model the statistic

$$T = \frac{y_n - \hat{y}_n^{(n-1)}}{\hat{\sigma}_{n-1}\sqrt{1 + x'_n(X'_{n-1}X_{n-1})^{-1}x_n}}$$

is $t_{n-K-1}$-distributed.

*Proof.* Recall a classical result: For independent random variables $X$ and $Y$ such that $X \sim N(0, 1)$ and $Y \sim \chi^2_n$, the statistic $\dfrac{X}{\sqrt{Y/n}}$ is $t_n$-distributed. In the proof of the theorem, we will show that statistic $T$ can be represented in the form mentioned above.

1) Note that for the random variable $y_n$, it holds that $y_n \sim N(x'_n\beta, \sigma^2)$.

For $\hat{y}_n^{(n-1)}$, we have that

$$\begin{aligned}
\hat{y}_n^{(n-1)} &= x'_n\big(X'_{n-1}X_{n-1}\big)^{-1}X'_{n-1}Y_{n-1} \\
&= x'_n\big(X'_{n-1}X_{n-1}\big)^{-1}X'_{n-1}\big(X_{n-1}\beta + \mathcal{E}_{n-1}\big) \\
&= x'_n\beta + x'_n\big(X'_{n-1}X_{n-1}\big)^{-1}X'_{n-1}\mathcal{E}_{n-1}.
\end{aligned}$$

This implies that $\hat{y}_n^{(n-1)}$ is a normally distributed random variable with parameters $\mathbb{E}\big(\hat{y}_n^{(n-1)}\big) = x'_n\beta$ and

$$\begin{aligned}
Var\big(\hat{y}_n^{(n-1)}\big) &= Var\big(x'_n\big(X'_{n-1}X_{n-1}\big)^{-1}X'_{n-1}\mathcal{E}_{n-1}\big) \\
&= x'_n(X'_{n-1}X_{n-1})^{-1}X'_{n-1}Var(\mathcal{E}_{n-1})X_{n-1}(X'_{n-1}X_{n-1})^{-1}x_n \\
&= \sigma^2 x'_n(X'_{n-1}X_{n-1})^{-1}x_n.
\end{aligned}$$

Since $y_n$ and $\hat{y}_n^{(n-1)}$ are independent and normally distributed, their difference is also normally distributed with parameters

$$\mathbb{E}(y_n - \hat{y}_n^{(n-1)}) = E(y_n) - E(\hat{y}_n^{(n-1)}) = x'_n\beta - x'_n\beta = 0,$$

$$Var(y_n - \hat{y}_n^{(n-1)}) = Var(y_n) + Var(\hat{y}_n^{(n-1)}) = \sigma^2(1 + x'_{n-1}(X'_{n-1}X_{n-1})^{-1}x_n).$$

Hence, the random variable

$$\frac{y_n - \hat{y}_n^{(n-1)}}{\sigma\sqrt{1 + x_n'(X_{n-1}'X_{n-1})^{-1}x_n}} \tag{18}$$

has standard normal distribution.

2) As shown in Appendix C.3.2 in [3], the statistic

$$\frac{(n-K-1)\hat{\sigma}_{n-1}^2}{\sigma^2} = \frac{(Y_{n-1} - \hat{Y}_{n-1})'(Y_{n-1} - \hat{Y}_{n-1})}{\sigma^2} \tag{19}$$

has $\chi^2$-distribution with $n-K-1$ degrees of freedom

3) It remains to show the independence of the statistics (18) and (19).

Note that $(n-K-1)\hat{\sigma}_{n-1}^2$ can be represented as the quadratic form applied to the standard normally distributed vector $\mathcal{E}_{n-1}$, since

$$\begin{aligned}
&(Y_{n-1} - \hat{Y}_{n-1})'(Y_{n-1} - \hat{Y}_{n-1}) \\
&\qquad = (Y_{n-1} - H_{n-1}Y_{n-1})'(Y_{n-1} - H_{n-1}Y_{n-1}) \\
&\qquad = Y_{n-1}'(I_{n-1} - H_{n-1})'(I_{n-1} - H_{n-1})Y_{n-1} \\
&\qquad = Y_{n-1}'(I_{n-1} - H_{n-1})Y_{n-1},
\end{aligned}$$

where we have applied the idempotence of the matrix $I_{n-1} - H_{n-1}$. Now, using the fact that $X_{n-1}(I_{n-1} - H_{n-1}) = 0$, we obtain that

$$\begin{aligned}
&Y_{n-1}'(I_{n-1} - H_{n-1})Y_{n-1} \\
&\qquad = (X_{n-1}\beta + \mathcal{E}_{n-1})'(I_{n-1} - H_{n-1})(X_{n-1}\beta + \mathcal{E}_{n-1}) \\
&\qquad = \mathcal{E}_{n-1}'(I_{n-1} - H_{n-1})\mathcal{E}_{n-1}.
\end{aligned}$$

At the same time, there holds

$$\hat{y}_n^{(n-1)} = x_n'\beta + x_n(X_{n-1}'X_{n-1})^{-1}X_{n-1}'\mathcal{E}_{n-1}.$$

It follows from Theorem 1.3.7 in [4] that for a standard normal random vector $\mathcal{E}$, the quadratic form $\mathcal{E}'A\mathcal{E}$ and linear transformation $B\mathcal{E}$ are independent if $BA = 0$. It is easy to see that this condition holds in our case above, since $I_{n-1} - H_{n-1}$ defines a matrix orthogonal to the columns of the matrix $X_{n-1}$. □

**Independence of the T-statistic and the summary statistic**

**Theorem 4.8.** In the Gauss linear model (14), the statistics

$$T_n = \frac{|y_n - \hat{y}_n^{(n-1)}|}{\hat{\sigma}_{n-1}\sqrt{1 + x_n'(X_{n-1}'X_{n-1})^{-1}x_n}}$$

and $S_n = (X_n, X_n'Y_n, Y_n'Y_n)$ are independent for every $n \geq 2$.

*Proof.* From Theorem 4.13, it follows that

$$T_n(X_n, Y_n) = \frac{\dfrac{e_n}{1-h_{nn}}}{\sqrt{\frac{1}{n-K-1}\Big(\sum_{i=1}^n e_i^2 - \dfrac{e_n^2}{1-h_{nn}}\Big)}\sqrt{1 + x_n'\Big(X_{n-1}'X_{n-1}\Big)^{-1}x_n}}.$$

As we can see, $T_n$ depends on $(X_n, Y_n)$ only through the values of the vector of residuals $E_n = (e_1, \ldots, e_n)' = Y_n - \hat{Y}_n^{(n)}$ and the fixed matrix $X_n$.

From this point on, we will omit the index $n$ in the notation.

It is also easy to see that $T(kE) = T(E)$, for every $k > 0$. Hence, for every $E$, we have that $T(E) = T(E/||E||)$, i.e., $T$ is defined by its values on the unit sphere in $[X]^{\perp}$.

On the other hand, for the statistic $S$, there holds

$$X'Y = X'P_XY \quad \text{and} \quad Y'Y = ||Y||^2 = ||E||^2 + ||P_XY||^2.$$

This implies that $S = g(X, P_XY, ||E||)$, where $g$ is a measurable function and $P_X$ is a projection matrix onto the subspace defined by the column vectors of the matrix $X$.

Recall that the vector of residuals $E$ is a projection of the standard normal vector $\mathcal{E}$ onto the subspace $[X]^{\perp}$, since

$$E = (I - P_X)Y = (I - P_X)(X\beta + \mathcal{E}) = (I - P_X)\mathcal{E}.$$

Therefore, we obtain

$$T = T\Big((I - P_X)\mathcal{E}/||(I - P_X)\mathcal{E}||\Big) \quad \text{and} \quad S = S(X, P_X\mathcal{E}, ||(I - P_X)\mathcal{E}||).$$

Note that $(I - P_X)\mathcal{E}$ and $P_X\mathcal{E}$ are independent by Corollary 1.2.4 in [4], since $P_X(I - P_X)' = 0$.

Also, $(I - P_X)\mathcal{E}/||(I - P_X)\mathcal{E}||$ and $||(I - P_X)\mathcal{E}||$ are independent by Lemma 4.10. Hence, we obtain that $T$ and $S$ are independent as measurable functions of independent random variables.

□

*Remark.* The proofs of the next results rely on the following facts.

1) If a random vector $X \in \mathbb{R}^n$ satisfies $X \sim MX$ for any orthogonal $n \times n$ matrix $M$, then $X$ is uniform on the unit sphere in $\mathbb{R}^n$, i.e., $X \sim U(S^{n-1})$.

2) Let $R \sim U(O(n))$, that is, $R$ is uniformly distributed on the collection of all orthogonal $n \times n$ matrices. If $x \in \mathbb{R}^n$ satisfies $||x|| = 1$, then $Rx \sim U(S^{n-1})$.

**Theorem 4.9.** Let $X \sim N_n(0, I_n)$. Define $\frac{X}{||X||}$ as

$$\frac{X}{||X||}(\omega) = \begin{cases} \frac{X(\omega)}{||X(\omega)||}, & \text{if } X(\omega) \neq 0 \\ Z(\omega), & \text{if } X(\omega) = 0 \end{cases}$$

where $Z \sim U(S^{n-1})$ and independent of $X$.

Then

1) $\frac{X}{||X||} \sim U(S^{n-1})$ and $||X||^2 \sim \chi^2_n$,

2) $\frac{X}{||X||}$ and $||X||$ are independent.

*Proof.* 1) (Distribution)

The fact that $||X||^2 \sim \chi^2_n$ follows from the definition of the $\chi^2$-distribution.

To show that $\frac{X}{||X||}$ is uniformly distributed on the unit sphere $S^{n-1}$ it is enough to show that $\frac{X}{||X||}$ is rotationally invariant (fact 1 in the remark). For every orthogonal $n \times n$ matrix $M$ it holds that

$$M\frac{X}{||X||} = \frac{MX}{||MX||} \sim \frac{X}{||X||},$$

since $X \sim MX$ for $X \sim N_n(0, I_n)$ and $||MX|| = ||X||$.

2) (Independence)

Let $R \sim U(O(n))$ and independent of $X$ and $Z$.

First, we show that $RX \sim X$. For every $A \in \mathcal{B}(\mathbb{R}^n)$ it holds that

$$\begin{aligned}\mathbb{P}(RX \in A) &= \mathbb{E}\mathbb{P}(RX \in A \mid R) \\ &= \mathbb{E}\mathbb{P}(X \in A \mid R) = \mathbb{P}(X \in A),\end{aligned}$$

since given any fixed orthogonal matrix $R_0$ the distribution of $R_0X$ is equal to $N(0, I_n)$.

To prove independence of $X/||X||$ and $||X||$ it is enough to prove independence of $RX/||X||$ and $||X||$.

For every $A \in \mathcal{B}(U(S^{n-1}))$ and $B \in \mathcal{B}([0, +\infty))$ it holds that

$$\begin{aligned}\mathbb{P}\big(\frac{RX}{||X||} \in A, ||X|| \in B\big) &= \mathbb{E}\big(\mathbb{1}_{\{RX/||X|| \in A\}}\mathbb{1}_{\{||X|| \in B\}}\big) \\ &= \mathbb{E}\mathbb{E}\big(\mathbb{1}_{\{RX/||X|| \in A\}}\mathbb{1}_{\{||X|| \in B\}} \mid X\big) = (\star).\end{aligned}$$

Given any value of $X$, the random variable $RX/||X||$ is $U(S^{n-1})$-distributed. Therefore, it holds that

$$\begin{aligned}(\star) = \mathbb{E}\Big(\mathbb{P}\big(RX/||X|| \in A\big)\,\mathbb{E}\big(\mathbb{1}_{\{||X|| \in B\}} \mid X\big)\Big) \\ = \mathbb{P}\big(RX/||X|| \in A\big)\mathbb{P}\big(||X|| \in B\big).\end{aligned}$$

□

**Lemma 4.10.** *Let $U$ be a subspace in $\mathbb{R}^n$ and $X \sim N(0, I_n)$. If $P$ is a projection matrix on $U$, then $PX/||PX||$ and $||PX||$ are independent.*

*Proof.* Let $m = \dim(U)$ and let $o_1, \ldots, o_m$ be an orthogonal basis of $U$. Then for the matrix $M = (o_1, \ldots, o_m) \in \mathbb{R}^{n \times m}$ it holds that

$$P = M(M'M)^{-1}M' = MM'.$$

Denote $M'X$ as $Z$. Then we have that $PX = MM'X = MZ$ and $Z \sim N(0, I_m)$.

Since $||PX|| = ||MZ|| = ||Z||$ it holds that

$$\begin{pmatrix} \frac{PX}{||PX||} \\ \\ ||PX|| \end{pmatrix} = \begin{pmatrix} \frac{MZ}{||MZ||} \\ \\ ||MZ|| \end{pmatrix} = \begin{pmatrix} M\frac{Z}{||Z||} \\ \\ ||Z|| \end{pmatrix}.$$

From Theorem 4.9 it follows that $Z/||Z||$ and $||Z||$ are independent. Therefore, $MZ/||Z||$ and $||Z||$ are also independent.

□

**Sherman-Morrison formula for Gauss linear model**

This appendix is a modified version of Appendices C.6 - C.8 in [3].

**Theorem 4.11.** (Sherman-Morrison formula) Let $A$ be an invertible $n \times n$ square matrix, let $u$ and $v$ be two column vectors in $\mathbb{R}^n$. Then, there holds the relation

$$\big(A + uv'\big)^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1 + v'A^{-1}u}. \tag{20}$$

**Lemma 4.12.** *Let $X_n$ be $n \times K$-matrix as in the definition of the Gauss linear model. Denote by $x_i'$ the $i$th row of $X_n$. Then, we have that*

$$\big(X_{n-1}'X_{n-1}\big)^{-1} = \big(X_n'X_n\big)^{-1} + \frac{(X_n'X_n)^{-1}x_nx_n'(X_n'X_n)'}{1 - x_n(X_n'X_n)^{-1}x_n'}.$$

*Proof.* The relation $X_n'X_n = \sum_{i=1}^n x_ix_i'$ implies that $X_{n-1}'X_{n-1} = X_n'X_n - x_nx_n'$. Now using the Sherman-Morrison formula, we get that

$$\big(X_{n-1}'X_{n-1}\big)^{-1} = \big(X_n'X_n - x_nx_n'\big)^{-1} = \big(X_n'X_n\big)^{-1} + \frac{(X_n'X_n)^{-1}x_nx_n'(X_n'X_n)'}{1 - x_n(X_n'X_n)^{-1}x_n'}.$$

□

**Theorem 4.13.** In the Gauss linear model (14), there hold the relations

$$1)\, y_n - \hat{y}_n^{(n-1)} = \frac{e_n}{1 - h_{nn}},$$

$$2)\, \hat{\sigma}_{n-1}^2 = \frac{1}{n - K - 1}\Big(\sum_{i=1}^n e_i^2 - \frac{e_n^2}{1 - h_{nn}}\Big),$$

where $E_n = (e_1, \ldots, e_n)'$ is a vector of residuals $Y_n - \hat{Y}_n^{(n)}$ and $h_{nn}$ is the $n$-th diagonal element of the hat-matrix $H = X_n(X_n'X_n)^{-1}X_n'$.

*Proof.* (Relation 1).

$$\begin{aligned}
y_n - \hat{y}_n^{(n-1)} &= y_n - x_n'(X_{n-1}'X_{n-1})^{-1}X_{n-1}'Y_{n-1} \\
&= y_n - x_n'\Big((X_n'X_n)^{-1} + \frac{(X_n'X_n)^{-1}x_nx_n'(X_n'X_n)^{-1}}{1 - x_n'(X_n'X_n)^{-1}x_n}\Big) \\
&\qquad \times \Big(X_n'Y_n - y_nx_n\Big) \\
&= \frac{1}{1-h_{nn}}\Big((1-h_{nn})y_n - (1-h_{nn})x_n'(X_n'X_n)^{-1}X_n'Y_n \\
&\qquad + (1-h_{nn})x_n'(X_n'X_n)^{-1}x_ny_n \\
&\qquad - x_n'(X_n'X_n)^{-1}x_nx_n'(X_n'X_n)^{-1}X_n'Y_n \\
&\qquad + x_n'(X_n'X_n)^{-1}x_nx_n'(X_n'X_n)^{-1}x_ny_n\Big) \\
&= \frac{1}{1-h_{nn}}\Big(y_n - h_{nn}y_n - \hat{y}_n^{(n)} + h_{nn}\hat{y}_n^{(n)} + h_{nn}y_n - h_{nn}^2y_n \\
&\qquad - h_{nn}\hat{y}_n^{(n)} + h_{nn}^2y_n\Big) = \frac{e_n}{1-h_{nn}}.
\end{aligned}$$

□

**Lemma 4.14.** *Let $\hat{\beta}_n$ be the least squares estimator for parameter $\beta$. Then*

$$\hat{\beta}_{n-1} = \hat{\beta}_n - \frac{(X_n'X_n)^{-1}x_ne_n}{1-h_{nn}}.$$

*Proof.* (Lemma)

The relation $X_{n-1}'Y_{n-1} = X_n'Y_n - y_nx_n$ and Lemma 4.12 imply that

$$\begin{aligned}
\hat{\beta}_{n-1} &= (X_{n-1}'X_{n-1})^{-1}X_{n-1}'Y_{n-1} \\
&= \Big((X_n'X_n)^{-1} + \frac{(X_n'X_n)^{-1}x_nx_n'(X_n'X_n)^{-1}}{1 - x_n'(X_n'X_n)^{-1}x_n}\Big)\Big(X_n'Y_n - y_nx_n\Big) \\
&= \frac{1}{1-h_{nn}}\Big(\big(1-h_{nn}\big)\big(\hat{\beta}_n - (X_n'X_n)^{-1}x_ny_n\big) \\
&\qquad + (X_n'X_n)^{-1}x_nx_n'(X_n'X_n)^{-1}X_n'Y_n \\
&\qquad - (X_n'X_n)^{-1}x_nx_n'(X_n'X_n)^{-1}x_ny_n\Big) \\
&= \hat{\beta}_n + \frac{1}{1-h_{nn}}\Big(h_{nn}(X_n'X_n)^{-1}x_ny_n \\
&\qquad - (X_n'X_n)^{-1}x_ny_n + (X_n'X_n)^{-1}x_n\hat{y}_n^{(n)} \\
&\qquad - h_{nn}(X_n'X_n)^{-1}x_ny_n\Big) \\
&= \hat{\beta}_n - \frac{(X_n'X_n)^{-1}x_ne_n}{1-h_{nn}}.
\end{aligned}$$

□

*Proof.* (Relation 2 in Theorem 4.13) It holds that

$$\begin{aligned}&\big(Y_{n-1} - X_{n-1}\hat{\beta}_{n-1}\big)'\big(Y_{n-1} - X_{n-1}\hat{\beta}_{n-1}\big)\\ &\qquad = \big(Y_n - X_n\hat{\beta}_{n-1}\big)'\big(Y_n - X_n\hat{\beta}_{n-1}\big) - \big(y_n - x_n'\hat{\beta}_{n-1}\big)^2.\end{aligned}$$

For the last term, we have that

$$\begin{aligned}\big(y_n - x_n'\hat{\beta}_{n-1}\big)^2 &= \Big(y_n - x_n'\Big(\hat{\beta}_n - \frac{(X_n'X_n)^{-1}x_n e_n}{1-h_{nn}}\Big)\Big)^2\\ &= \Big(y_n - x_n'\hat{\beta}_n + \frac{x_n'(X_n'X_n)^{-1}x_n e_n}{1-h_{nn}}\Big)^2\\ &= \Big(e_n + \frac{h_{nn}\, e_n}{1-h_{nn}}\Big)^2 = \frac{e_n^2}{(1-h_{nn})^2} = (**).\end{aligned}$$

For the first term, it holds that

$$\begin{aligned}&\big(Y_n - X_n\hat{\beta}_{n-1}\big)'\big(Y_n - X_n\hat{\beta}_{n-1}\big)\\ &\quad = \Big(Y_n - X_n\Big(\hat{\beta}_n - \frac{(X_n'X_n)^{-1}x_n e_n}{1-h_{nn}}\Big)\Big)'\Big(Y_n - X_n\Big(\hat{\beta}_n - \frac{(X_n'X_n)^{-1}x_n e_n}{1-h_{nn}}\Big)\Big)\\ &\quad = \Big(E_n + \frac{X_n(X_n'X_n)^{-1}x_n e_n}{1-h_{nn}}\Big)'\Big(E_n + \frac{X_n(X_n'X_n)^{-1}x_n e_n}{1-h_{nn}}\Big)\\ &\quad = \Big(E_n + \frac{e_n h_{\bullet n}}{1-h_{nn}}\Big)'\Big(E_n + \frac{e_n h_{\bullet n}}{1-h_{nn}}\Big)\\ &\quad = E_n'E_n + \frac{2e_n}{1-h_{nn}}h_{\bullet n}'E_n + \Big(\frac{e_n}{1-h_{nn}}\Big)^2 h_{\bullet n}'h_{\bullet n} = (*),\end{aligned}$$

where $h_{\bullet n}$ is the $n$-th column of the hat-matrix $H$.

Since the vector of residuals $E_n$ is orthogonal to the column space $[X]$, we have that $H'E_n = 0$. Hence, $h_{\bullet n}'E_n = 0$.

The matrix $H$ is symmetric and idempotent (i.e. $HH = H$). Therefore,

$$h_{\bullet n}'h_{\bullet n} = h_{n\bullet}h_{\bullet n} = h_{nn}.$$

These two arguments imply that

$$(*) = E_n'E_n + \frac{h_{nn}\, e_n^2}{(1-h_{nn})^2}.$$

Finally, combining $(*)$ and $(**)$, we obtain

$$(n-K-1)\sigma_{n-1}^2 = E_n'E_n + \frac{h_{nn}\, e_n^2}{(1-h_{nn})^2} - \frac{e_n^2}{(1-h_{nn})^2} = \sum_{i=1}^{n} e_i^2 - \frac{e_n^2}{1-h_{nn}}.$$

□

## 4.3 MVA model

In this section, we discuss the results presented in Section 7 of [2] and provide rigorous proofs.

The last model in the overview of applications is the so-called MVA model, which combines assumptions from the IID model and the Gauss linear model with an additional assumption about the normal distribution of regressors. More formally, for random vectors $(\tilde{x}_n, y_n) \in \mathbb{R}^K \times \mathbb{R},\ n \geq 1$, we have that

$$y_n = \alpha + \tilde{x}_n'\beta + \varepsilon_n$$

where for all $n \geq 1$, it holds that

1. vectors $\tilde{x}_n$ are independent and $N_K(m, \Sigma)$-distributed.
2. $\varepsilon_n$ are independent and $N(0, \sigma^2)$-distributed random variables.
3. $\tilde{x}_n'$ and $\varepsilon_n$ are independent.

Throughout this section, we will use the following notation

$$Y_n = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \tilde{X}_n = \begin{bmatrix} \tilde{x}_1' \\ \tilde{x}_2' \\ \vdots \\ \tilde{x}_n' \end{bmatrix}, \quad \mathcal{E}_n = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Notice that $\tilde{X}_n$ is a matrix of stochastic regressors. Also, denote

$$X_n = \begin{bmatrix} 1 & \tilde{x}_1' \\ 1 & \tilde{x}_2' \\ & \vdots \\ 1 & \tilde{x}_n' \end{bmatrix} \quad \text{and} \quad i_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Denote as $x_n'$ the row vector $(1,\ \tilde{x}_n')$ for all $n \geq 1$.

*Remark.* (on reducing the number of regressors in the MVA model)

Consider the following two properties of the MVA model:

$$\mathbb{E}(y_n \mid x_n) \text{ is linear in } x_n. \tag{21}$$

$$Var(y_n \mid x_n) \text{ is constant.} \tag{22}$$

It is worth mentioning that in the MVA model, if we believe that some regressors are less important than others, then we can remove the less important regressors and the properties (21) and (22) will still hold.

The crucial property of the MVA model which allows such a procedure is that the joint distribution of regressors and errors is Gaussian. As we can see in the next lemma (Proposition 3.13 in [5]), for Gaussian vectors, there hold the properties equivalent to (21) and (22).

**Lemma 4.15.** *If $X_1$ and $X_2$ are two (Gaussian) subvectors of a Gaussian vector $X$, then*

- $\mathbb{E}(X_1 \mid X_2) = a + b'X_2$ *and*
- $Var(X_1 \mid X_2) = \Sigma$ *(constant matrix independent of $X_2$)*

*for a vector $a$ and matrices $b$ and $\Sigma$ of appropriate dimension.*

**Theorem 4.16.** If $x_{+,n}$ and $x_{-,n}$ are the vectors of important and unimportant regressors respectively and $\beta_+$ and $\beta_-$ are the vectors of corresponding coordinates of $\beta$, then

1. $\mathbb{E}(y_n \mid x_{+,n})$ is linear in $x_{+,n}$.
2. $Var(y_n \mid x_{+,n})$ is constant.

*Proof.* Note that since the vectors $x'_{+,n}$ and $x'_{-,n}$ are disjoint and together form the vector $\tilde{x}'_n$, we have that $\tilde{x}'_n\beta = x'_{+,n}\beta_+ + x'_{-,n}\beta_-$.

1. Using the representation above, we get

$$\begin{aligned}\mathbb{E}(y_n \mid x_{+,n}) &= \mathbb{E}(\alpha + \tilde{x}'_n\beta + \varepsilon_n \mid x_{+,n}) \\ &= \mathbb{E}(\alpha + x'_{+,n}\beta_+ + x'_{-,n}\beta_- + \varepsilon_n \mid x_{+,n}) \\ &= \mathbb{E}(\alpha + x'_{+,n}\beta_+ + \varepsilon_n \mid x_{+,n}) + \mathbb{E}(x_{-,n}\beta_- \mid x_{+,n}) \\ &= \alpha + x'_{+,n}\beta_+ + \mathbb{E}(x_{-,n}\beta_- \mid x_{+,n}).\end{aligned}$$

Since the vector $(x_{-,n}\beta_-,\, x_{+,n})$ is Gaussian, using Lemma 4.15, we obtain

$$\mathbb{E}(x'_{-,n}\beta_- \mid x_{+,n}) = \alpha_0 + x'_{+,n}\beta_0$$

for some constant $\alpha_0$ and $\beta_0$ and, finally

$$\mathbb{E}(y_n \mid x_{+,n}) = (\alpha + \alpha_0) + x'_{+,n}(\beta + \beta_0).$$

2. In this case, we will use the following properties of conditional variance: For a random variable $Y$ and $\sigma$-algebra $\mathcal{C}$ there hold the relations

$$Var(a + Y \mid \mathcal{C}) = Var(Y \mid \mathcal{C}) \text{ for any constant } a \in \mathbb{R}.$$

$$Var(X + Y \mid \mathcal{C}) = Var(Y \mid \mathcal{C}) \text{ for any } \mathcal{C}\text{-measurable random variable } X.$$

$$Var(Y + \mathcal{E} \mid \mathcal{C}) = Var(Y \mid \mathcal{C}) + Var(\mathcal{E}) \text{ for any random variable } \mathcal{E}$$
$$\text{independent of } \mathcal{C} \text{ and } Y.$$

These statements imply that

$$\begin{aligned}Var(y_n \mid x_{+,n}) &= Var(\alpha + x'_{+,n}\beta_+ + x'_{-,n}\beta_- + \varepsilon_n \mid x_{+,n}) \\ &= Var(\varepsilon_n) + Var(x'_{-,n}\beta_- \mid x_{+,n}) = \sigma^2 + \delta^2.\end{aligned}$$

□

To construct conformal intervals, we start by defining an ATTS sequence of statistics for the considered model.

**Lemma 4.17.** *The sequence of statistics $S_n, n \geq 1$, such that*

$$S_n(X_n, Y_n) = (X_n'X_n, X_n'Y_n, Y_n'Y_n) \tag{23}$$

*is ATTS for the MVA model.*

*Proof.* The algebraic transitivity of $S_n$ follows from the relations

$$X_n'X_n = \sum_{i=1}^{n} x_i x_i', \;\; X_n'Y_n = \sum_{i=1}^{n} y_i x_i \text{ and } Y_n'Y_n = \sum_{i=1}^{n} y_i^2,$$

which imply that

$$X_n'X_n = X_{n-1}'X_{n-1} + x_n x_n' \text{ and } X_n'Y_n = X_{n-1}'Y_{n-1} + y_n x_n.$$

To prove sufficiency, we apply the Neyman-Fisher factorization theorem. First note that $y_i \mid \tilde{x}_i \sim N(\alpha + \tilde{x}_i'\beta, \sigma^2)$. Now, since for every $i$, the joint density $p(\tilde{x}_i, y_i) = p_{y|\tilde{x}}(y_i \mid \tilde{x}_i)\, p_{\tilde{x}}(\tilde{x}_i)$, we have that

$$\begin{aligned}
L\big((\tilde{x}_1, y_1), \ldots, (\tilde{x}_n, y_n)\big) &= \prod_{i=1}^{n} p(\tilde{x}_i, y_i) = \prod_{i=1}^{n} p_{y|\tilde{x}}(y_i \mid \tilde{x}_i)\, p_{\tilde{x}}(\tilde{x}_i) \\
&= (2\pi\sigma^2)^{-n/2} \exp\Big( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \tilde{x}_i'\beta)^2 \Big) \\
&\qquad \times \det(2\pi\Sigma)^{-1/2} \exp\Big( -\frac{1}{2} \sum_{i=1}^{n} (\tilde{x}_i - m)'\Sigma^{-1}(\tilde{x}_i - m) \Big) \\
&= C \exp\Big( -\frac{1}{2\sigma^2} \big(Y_n - \alpha i_n - \tilde{X}_n\beta\big)'\big(Y_n - \alpha i_n - \tilde{X}_n\beta\big) \Big) \\
&\qquad \times \exp\Big( -\frac{1}{2} \sum_{i=1}^{n} \big(\tilde{x}_i'\Sigma^{-1}\tilde{x}_i - 2m'\Sigma^{-1}\tilde{x}_i + m'\Sigma^{-1}m\big) \Big).
\end{aligned}$$

For the first factor, note that

$$\begin{aligned}
&\big(Y_n - \alpha i_n - \tilde{X}_n\beta\big)'\big(Y_n - \alpha i_n - \tilde{X}_n\beta\big) \\
&\qquad = Y_n'Y_n - 2\beta'\tilde{X}_n'Y_n + \beta'\tilde{X}_n'\tilde{X}_n\beta - 2\alpha Y_n'i_n + 2\alpha\beta'\tilde{X}_n'i_n + \alpha^2 i_n'i_n,
\end{aligned}$$

which clearly depends on the sample through the values of $S_n$. For the second factor, we have

$$\begin{aligned}
\sum_{i=1}^{n} \tilde{x}_i\Sigma^{-1}\tilde{x}_i = tr\Big(\sum_{i=1}^{n} \tilde{x}_i'\Sigma^{-1}\tilde{x}_i\Big) &= \sum_{i=1}^{n} tr\big(\tilde{x}_i'\Sigma^{-1}\tilde{x}_i\big) \\
&= \sum_{i=1}^{n} tr\big(\Sigma^{-1}\tilde{x}_i\tilde{x}_i'\big) = tr\big(\Sigma^{-1}\tilde{X}_n'\tilde{X}_n\big),
\end{aligned}$$

$$\sum_{i=1}^{n} m'\Sigma^{-1}\tilde{x}_i = m'\Sigma^{-1}\sum_{i=1}^{n} \tilde{x}_i = m'\Sigma^{-1}\tilde{X}_n'i_n$$

and we also obtain the dependence on the sample through the values of $S_n$. □

To define a nonconformity measure for the MVA model, we first fit a line to the data points $(x_1, y_1), \ldots, (x_n, y_n)$ using the ridge regression with parameter $a > 0$ and consider the corresponding vector of residuals

$$E_n = (e_1, \ldots, e_n)' = Y_n - X_n\big(X_n'X_n + aI_n\big)^{-1}X_n'Y_n. \tag{24}$$

Now the nonconformity measure will be defined as follows

$$A_n\big((z_1, \ldots, z_{n-1}), z_n\big) = \sqrt{\frac{n-1}{n}}\,\frac{e_n - \overline{e}_{n-1}}{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n-1}(e_i - \overline{e}_{n-1})^2}} \tag{25}$$

*Remark.* The measure in (25) is indeed a function of the summary statistic $S_{n-1}$ and the new observation $z_n = (x_n, y_n)$, as required in the definition of nonconformity measures.

*Proof.* From the definition of residuals (24), we have

$$X_n'X_n = X_{n-1}'X_{n-1} + x_nx_n' \text{ and } X_n'Y_n = X_{n-1}'Y_{n-1} + y_nx_n,$$

which implies that $\big(X_n'X_n + aI_n\big)^{-1}X_n'Y_n = C\big(S_{n-1}, (x_n, y_n)\big)$. From this, we obtain

$$e_n - \overline{e}_{n-1} = y_n - x_n'C - \frac{1}{n-1}i_{n-1}'\big(Y_{n-1} - X_{n-1}C\big)$$

and

$$\sum_{i=1}^{n-1}\big(e_i - \overline{e}_{n-1}\big)^2 = \sum_{i=1}^{n-1}e_i^2 - (n-1)\,\overline{e}_{n-1}^2.$$

As seen above, $\overline{e}_{n-1}$ depends on $S_{n-1}$ and $(x_n, y_n)$ . Also, we have that

$$\begin{aligned}\sum_{i=1}^{n-1}e_i^2 &= \big(Y_{n-1} - X_{n-1}C\big)'\big(Y_{n-1} - X_{n-1}C\big)' \\ &= Y_{n-1}'Y_{n-1} - 2C'X_{n-1}'Y_{n-1} + C'X_{n-1}'X_{n-1}C,\end{aligned}$$

where all the terms are functions of $S_{n-1}$ and $(x_n, y_n)$.

□

Recall that in the construction of conformal intervals, we rely on the conditional distribution of a nonconformity measure $A_n$ given the summary statistic $S_n$. As we can see from Theorem 4.22, under the assumptions of the MVA model, the nonconformity measure (25) is independent of the summary statistic (23) and has $t_{n-2}$-distribution. This result implies that for the conformal interval $\Gamma_n^\varepsilon$,

$$\Gamma_n^\varepsilon = \big\{\, y \in \mathbb{R} \mid p_n(y) \geq \varepsilon \,\big\} = \big\{\, y \in \mathbb{R} \mid \mathbb{P}\big(A_n \geq A_n(y) \mid S_n = S_n(y)\big) \geq \varepsilon \,\big\}$$

$$= \big\{\, y \in \mathbb{R} \mid \mathbb{P}\big(A_n \geq A_n(y)\big) \geq \varepsilon \,\big\} = \big\{\, y \in \mathbb{R} \mid A_n(y) \geq |t_{n-2}^{\varepsilon/2}| \,\big\}$$

$$= \Big\{y \in \mathbb{R} \mid \sqrt{\frac{n-1}{n}}\,\frac{|\,e_n - \overline{e}_{n-1}\,|}{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n-1}(e_i - \overline{e}_{n-1})^2}} \geq |\,t_{n-2}^{\varepsilon/2}\,|\Big\}.$$

### 4.3.1 Appendix

We will omit the index $n$ in the notation if the context is clear.

**Theorem 4.18.** In the MVA model, consider the vector of residuals for ridge regression:

$$E_n = Y_n - X_n(X_n'X_n + aI_n)^{-1}X_n'Y_n.$$

Then the following statistic, which is defined through the vector $E_n$,

$$A = A(E_n) = \sqrt{\frac{n-1}{n}}\frac{e_n - \overline{e}_{n-1}}{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n-1}(e_i - \overline{e}_{n-1})^2}}$$

has $t_{n-2}$ distribution.

*Proof.* It holds that

$$\begin{aligned} e_n - \overline{e}_{n-1} &= e_n - \frac{e_1 + \cdots + e_{n-1}}{n-1} = \frac{(n-1)e_n - (e_1 + \cdots + e_{n-1})}{n-1} \\ &= \frac{ne_n - n\overline{e}_n}{n-1} = \frac{n}{n-1}(e_n - \overline{e}_n) \end{aligned}$$

and

$$\begin{aligned} \sum_{i=1}^{n-1}(e_i - e_{n-1})^2 &= \sum_{i=1}^{n-1}\left(e_i - \overline{e}_n + \overline{e}_n - \overline{e}_{n-1}\right)^2 \\ &= \sum_{i=1}^{n-1}\left((e_i - \overline{e}_n)^2 + 2(e_i - \overline{e}_n)(\overline{e}_n - \overline{e}_{n-1}) + (\overline{e}_n - \overline{e}_{n-1})^2\right) \\ &= \sum_{i=1}^{n-1}(e_i - \overline{e}_n)^2 + 2(\overline{e}_n - \overline{e}_{n-1})\sum_{i=1}^{n-1}(e_i - \overline{e}_n) + \sum_{i=1}^{n-1}(\overline{e}_n - \overline{e}_{n-1})^2 \\ &= \sum_{i=1}^{n-1}(e_i - \overline{e}_n)^2 + 2(\overline{e}_n - \overline{e}_{n-1})(n-1)(\overline{e}_{n-1} - \overline{e}_n) + (n-1)(\overline{e}_n - \overline{e}_{n-1})^2 \\ &= \sum_{i=1}^{n-1}(e_i - \overline{e}_n)^2 - (n-1)(\overline{e}_n - \overline{e}_{n-1})^2 = (\star). \end{aligned}$$

Since $\overline{e}_n - \overline{e}_{n-1} = (e_n - \overline{e}_n)/(n-1)$, it holds that

$$(\star) = \sum_{i=1}^{n}(e_i - \overline{e}_n)^2 - (e_n - \overline{e}_n)^2 - \frac{(e_n - \overline{e}_n)^2}{n-1} = \sum_{i=1}^{n}(e_i - \overline{e}_n)^2 - \frac{n}{n-1}(e_n - \overline{e}_n)^2.$$

From the computations above, it follows that

$$\begin{aligned} A(E_n) &= \sqrt{\frac{n-1}{n}}\frac{e_n - \overline{e}_{n-1}}{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n-1}(e_i - \overline{e}_{n-1})^2}} \\ &= \sqrt{\frac{n-1}{n}}\frac{\frac{n}{n-1}(e_n - \overline{e}_n)}{\sqrt{\frac{1}{n-2}\Big(\sum_{i=1}^{n}(e_i - \overline{e}_n)^2 - \frac{n}{n-1}(e_n - \overline{e}_n)^2\Big)}}. \end{aligned}$$

Note that the values $e_i - \overline{e}_n$ are coordinates of the projection of the vector $E_n$ onto the subspace $[i_n]^\perp$, which is equal to the vector $E_n - P_{i_n}E_n$. As we have already seen in the section on the Gauss linear model, the statistic $A$ can be represented as

$$A = g\Big(\frac{E_n - P_{i_n}E_n}{||E_n - P_{i_n}E_n||}\Big)$$

where $g$ is a measurable function.

Previously we have also proved that in the case of a standard normally distributed vector $Z$, the statistic $g\Big(\frac{Z_n - P_{i_n}Z_n}{||Z_n - P_{i_n}Z_n||}\Big)$ has $t_{n-2}$ distribution. An important property of the vector $\dfrac{Z_n - P_{i_n}Z_n}{||Z_n - P_{i_n}Z_n||}$ is that it is uniformly distributed on a unit sphere in the subspace $[i_n]^\perp$. Hence, it is enough to show that the vector $\dfrac{E_n - P_{i_n}E_n}{||E_n - P_{i_n}E_n||}$ is also uniformly distributed on the unit sphere in $[i_n]^\perp$, which we proved in Lemma 4.21. □

**Lemma 4.19.** *Let $R$ be an orthogonal matrix that leaves the subspace $[i_n]$ unchanged. Then under the assumptions of the MVA model, it holds that $(Y, X)' \sim (RY, RX)'$.*

*Proof.* Recall that the matrix $X$ consists of a column vector $i_n = [1, \dots, 1]'$ and a matrix $\tilde{X}$ of $n$ i.i.d. normally distributed random row vectors $\tilde{x}_1', \dots, \tilde{x}_n'$, i.e., $X = [i_n \mid \tilde{X}]$.

Since the vectors $\tilde{x}_1', \dots, \tilde{x}_n'$ have the same multivariate normal distribution with parameters $m = (m_1, \dots, m_K)'$ and $\Sigma$, we have for every $i = 1, \dots, n$ that $\tilde{x}_i' = m' + u_i' \Sigma^{1/2}$, where the random vectors $u_i'$ are independent and $N_K(0, \sigma^2 I_K)$-distributed. The latter implies that

$$\tilde{X} = \begin{bmatrix} \tilde{x}_1' \\ \vdots \\ \tilde{x}_n' \end{bmatrix} = \begin{bmatrix} m' + u_1' \, \Sigma^{1/2} \\ \vdots \\ m' + u_n' \, \Sigma^{1/2} \end{bmatrix} = \begin{bmatrix} m' \\ \vdots \\ m' \end{bmatrix} + \begin{bmatrix} u_1' \\ \vdots \\ u_n' \end{bmatrix} \Sigma^{1/2}$$

$$= \begin{bmatrix} m_1 i_n \mid & \dots & \mid m_K i_n \end{bmatrix} + \begin{bmatrix} u_{\bullet 1} \mid & \dots & \mid u_{\bullet K} \end{bmatrix} \Sigma^{1/2},$$

where $u_{\bullet i}$ is $N_n(0, \sigma^2 I_n)$-distributed for all $i = 1, \dots, n$. Now we have

$$R\tilde{X} = \begin{pmatrix} m_1 R\, i_n \mid & \dots & \mid m_K R\, i_n \end{pmatrix} + \begin{pmatrix} R\, u_{\bullet 1} \mid & \dots & \mid R\, u_{\bullet K} \end{pmatrix} \Sigma^{1/2}.$$

Since $R\, i_n = i_n$ and, for every $N_n(0, \sigma^2 I_n)$-distributed vector $U$ there holds $RU \sim U$, we have that $R\tilde{X} \sim \tilde{X}$ and, therefore, $RX \sim X$.

Now we know that $X$ and $\mathcal{E}$ are independent. Since $RX \sim X$ and $R\mathcal{E} \sim \mathcal{E}$, we also have that $(X, \mathcal{E})' \sim (RX, R\mathcal{E})'$. It follows that

$$\begin{pmatrix} Y \\ X \end{pmatrix} = \begin{pmatrix} \alpha i_n + \tilde{X}\beta + \mathcal{E} \\ X \end{pmatrix} \sim \begin{pmatrix} R\alpha i_n + R\tilde{X}\beta + R\mathcal{E} \\ RX \end{pmatrix} = \begin{pmatrix} RY \\ RX \end{pmatrix}.$$

□

**Lemma 4.20.** *In the MVA model, consider the vector of residuals*

$$E = Y - X(X'X + aI)^{-1}X'Y.$$

*Then for the orthogonal matrix $R$ defined above, there holds $RE \sim E$.*

*Proof.* We have that $E = r(Y, X)$, where $r$ is a measurable function. Now it holds that

$$\begin{aligned} RE &= RY - RX(X'X + aI)^{-1}X'Y \\ &= RY - RX(X'R'RX + aI)^{-1}X'R'RY \\ &= r(RY, RX). \end{aligned}$$

Since $r$ is measurable and, as we have shown in Lemma 4.19, $(Y, X)' \sim (RY, RX)'$, we obtain

$$E = r(Y, X) \sim r(RY, RX) = RE.$$

□

**Lemma 4.21.** *For the orthogonal matrix $R$ defined above and the vector $E$ in the MVA model, it holds that*

$$R\,\frac{E - P_{i_n}E}{||E - P_{i_n}E||} \sim \frac{E - P_{i_n}E}{||E - P_{i_n}E||},$$

*i.e., the invariance of the distribution with respect to the rotation around the vector $i_n$.*

*Proof.* Let $\dfrac{E - P_{i_n}E}{||E - P_{i_n}E||} = g(E)$, where $g$ is a measurable function. Then

$$R\,\frac{E - P_{i_n}E}{||E - P_{i_n}E||} = \frac{RE - RP_{i_n}E}{||RE - RP_{i_n}E||} = \frac{RE - P_{i_n}RE}{||RE - P_{i_n}RE||} = g(RE),$$

where we have used that $P_{i_n}R = RP_{i_n}$ (see Lemma 4.23) and that the orthogonal matrix $R$ does not change the length of a vector. Finally, $RE \sim E$ implies $g(E) \sim g(RE)$.

□

**Theorem 4.22.** In the MVA model, conditionally on the statistic $(X'X, X'Y, Y'Y)$, the nonconformity measure (25) is $t_{n-2}$-distributed.

*Proof.* First we show that for an orthogonal matrix $R$ that leaves vectors of $[i_n]^\perp$ unchanged, it holds that the joint distribution of $(X, Y)$ and $(X'X, X'Y, Y'Y)$ is equal to the joint distribution of $(RX, RY)$ and $(X'X, X'Y, Y'Y)$.

As we already proved in Lemma 4.19, $(RX, RY) \sim (X, Y)$. This implies that

$$\begin{bmatrix} (X, Y) \\ (X'X, X'Y, Y'Y) \end{bmatrix} \sim \begin{bmatrix} (RX, RY) \\ ((RX)'RX, (RX)'RY, (RY)'RY) \end{bmatrix}.$$

Since $R'R = I$, we have that

$$\begin{bmatrix} (RX, RY) \\ \big((RX)'RX, (RX)'RY, (RY)'RY\big) \end{bmatrix} = \begin{bmatrix} (RX, RY) \\ \big(X'X, X'Y, Y'Y\big) \end{bmatrix}.$$

From this, it follows that, given $(X'X, X'Y, Y'Y)$, the conditional distributions of $(X, Y)$ and $(RX, RY)$ are equal. Now using the same computations as in Lemma 4.20 we get that, given $(X'X, X'Y, Y'Y)$, the conditional distributions of the random vectors $RE$ and $E$ are also equal. By the analogy with Lemma 4.21 the latter implies that conditionally on $(X'X, X'Y, Y'Y)$ the vector $\dfrac{E - P_{i_n}E}{||E - P_{i_n}E||}$ is uniformly distributed on the unit sphere in $[i_n]^\perp$ and therefore the statistic $g\Big(\dfrac{E - P_{i_n}E}{||E - P_{i_n}E||}\Big)$ has $t_{n-2}$-distribution given $(X'X, X'Y, Y'Y)$.

□

**Lemma 4.23.** *If $P$ is a projection matrix onto a subspace $U$ of $\mathbb{R}^n$ and $R$ is an orthogonal matrix which leaves vectors of $U^\perp$ unchanged, then $PR = RP$.*

*Proof.* For any vector $v \in U^\perp$, it holds that $Rv = v$. For the projection matrix $P$ and any vectors $u \in U$, $v \in U^\perp$, we have that $Pu = u$ and $Pv = 0$.

Since every vector $x \in \mathbb{R}^n$ can be represented as $x = x_U + x_{U^\perp}$, where $x_U \in U$ and $x_{U^\perp} \in U^\perp$, we obtain

$$PRx = PR(x_U + x_{U^\perp}) = PRx_U + PRx_{U^\perp} = PRx_U = Rx_U,$$
$$RPx = RP(x_U + x_{U^\perp}) = RPx_U + RPx_{U^\perp} = Rx_U.$$

□

# 5 Conclusions

In this thesis, we have given an overview of conformal prediction, which is a method of constructing prediction intervals, and discussed its applications to different regression models. We mainly followed the paper "Online predictive linear regression" by Vovk, Nouretdinov, and Gammerman (see [2]). We have considered its main results in detail and provided rigorous proofs of some statements.

Using an illustrative example of the exchangeability model, we have presented and discussed the main concepts of conformal prediction, i.e., sufficient statistic, nonconformity measure, and randomized $p$-values. For the exchangeability model, we have also considered a simplified version of the theorem about the (conservative) validity of conformal intervals and introduced the idea of randomized conformal intervals.

In Section 3, we have provided an overview of the concept of conformal prediction its two key properties: strong validity and optimality. For completeness of the presentation, we have discussed the classical notions of deterministic and randomized $p$-values of a random variable and proved a result on the distribution of randomized $p$-value, which is crucial for constructing conformal intervals.

In Section 4, we have considered applications of conformal prediction to three different regression models, i.e., the IID model, the Gauss linear model, and the MVA model. For the IID model, we have presented the concept of a bag and proved several of its measure-theoretical properties. For the Gauss linear model and for the MVA model, we have shown that their previously known prediction intervals can also be obtained as conformal intervals. To prove the latter facts, we have discussed some theorems about the independence of the considered nonconformity measures and ATTS statistics for the corresponding models. In the proofs, we mainly showed the conditional and unconditional versions of rotational invariance of the nonconformity measures.

# References


[1] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Springer, New York, 2005.

[2] V. Vovk, I. Nouretdinov, and A. Gammerman, "Online predictive linear regression," *The Annals of Statistics*, vol. 37, no. 3, p. 1566–1590, 2009.

[3] D. Montgomery, E. Peck, and G. Vining, *Introduction to Linear Regression Analysis*. Wiley, New York, 2001.

[4] R. Christensen, *Plane Answers to Complex Questions – The Theory of Linear Models*. Springer, New York, 2002.

[5] M. Eaton, *Multivariate Statistics: A Vector Space Approach*. Wiley, New York, 1983.