Comparing Between the Imported and Local Bottled Drinking Water by LASSO Regression

Ali Sadig Mohommed Bager a, A Hussein b,c

aAl-Muthanna University, Al-Muthanna, Samawah, Iraq
bDepartment of Civil Engineering, College of Engineering, Al-Muthanna University
Al-Muthanna, Samawah, Iraq.
cCivil Engineering Research Group, School of Computing, Science and Engineering, University of Salford, Salford, Greater Manchester M5 4WT, United Kingdom.

a.hussein3@edu.salford.ac.uk, amjad.muhamad@mu.edu.iq

Abstract. Predict the significant variables of the quality measurement results of 10 and 5 kinds of imported and local bottled drinking water, respectively, tested in the Samawah city, Iraq, using regression analysis (LASSO). These variables were pH, turbidity, total dissolved solids, total hardness, calcium, magnesium, sodium, fluoride, nitrates, sulphates, chlorides, iron, manganese. Ph was selected as a y dependant, and others were chosen as x independents. The results showed that nitrates, sulphates, and magnesium were insignificant (Beta > 0.05) in imported and local bottled drinking water, while sodium was insignificant (Beta > 0.05) in local bottled drinking water only. From these results, LASSO regression gave better results.

Keywords: Environment; bottled drinking water packaging; local brands; nitrates; LASSO regression.

1. Introduction

The first time the water was packaged in a bottle to use as bottled drinking water was in 1621, UK [1]. Most people prefer bottled drinking water to tap water because of taste and water quality’s regularity over time [2,3]. Another reason was tap water through flow from the source until the consumer may change due to the materials they are exposed from the surrounding environment. In contrast, it is very low in bottled drinking water because it is placed in sealed packaging [4].

In Iraq, the packaged drinking water industry has developed rapidly in the past 30 years and has a high production capacity. Bottled water has different sizes, most of which are used once, and items imported from other countries can be obtained on bottled water and are available on the local market.

Lasso (Least Absolute Shrinkage and Selection Operator) is a linear model estimation method proposed by Tibshirani [5]. It refers to a set of processes that use L1 penalty points to narrow parameter estimates and perform automatic variable selection: the L1 penalty at least squares regression. Like garrote, it shrinks some coefficients while setting the remaining coefficients to precisely zero. Tibshirani believes that lasso is better than ordinary least squares (OLS) regression for two reasons: First of all, the over-specified OLS model usually has a small deviation but a large
variance, which is not conducive to its prediction accuracy. This effect can be improved by reducing or setting individual coefficients to zero and swapping some deviations to minimize model variance. Secondly, OLS models may sometimes have a large number of small coefficients, which add little value to the model and complicate the interpretation of effects [6]. The research aims to find the insignificant measurements from all chemical and physical measures tested on the imported and local bottled drinking water by Hussein and Mohammed [4]. LASSO regression will use to assess this aim.

2. Methodology
The methodology involved two sections in this research: sample collection and data analysis (LASSO regression).

2.1. Samples Collection
All the samples that will be analyzed by LASSO regression, as shown in Tables 1 and 2, had been tested according to APHA., (1995) [7,8,9].

Table 1. The quality measurement results of drinking water for the local items [4].

| Symbol | Mn | Fe | Cl | No | F | Na | So |
|--------|----|----|----|----|---|----|----|
|        | mg/l|    |    |    |   |    |    |
| L1     | 0.00| 0.010 | 20 | 4.10| 0.600 | 20| 31 |
| L2     | 0.00| 0.000 | 36 | 9   | 0.095 | 49.4| 47 |
| L3     | 0.00| 0.015 | 187| 4.30| 1.100 | 50 | 210|
| L4     | 0.00| 0.000 | 16 | 0.115| 0.300 | 13 | 48 |
| L5     | 0.01| 0.010 | 197| 1.100| 1.150 | 22.5| 48 |
| KSA Standard | 0.05| 0.300 | 250| 10  | 0.6-1 | - | 250|

| Element | Mg | CaCO3 | Ca | TDS | TU | pH |
|---------|----|-------|----|-----|----|----|
|         | mg/l|       |    |     |    |    |
| L1      | 5.05| 46 | 10 | 118 | 0.28 | 7.2 |
| L2      | 3.60| 184 | 68 | 123 | 0.22 | 7.5 |
| L3      | 6.10| 145 | 48 | 162 | 0.24 | 7.6 |
| L4      | 2.35| 20 | 4.1 | 19.8 | 0.10 | 7.3 |
| L5      | 3.60| 130 | 46 | 204 | 0.21 | 7.1 |
| KSA Standard | 30.0| 300 | 75 | 100-700 | 5.00 | 6.5-8.5 |

*Local, b the Kingdom of Saudi Arabia, 1Manganese, 2Iron, 3Chloride, 4Nitrate, 5Fluoride, 6Sodium, 7Sulphates, 8Magnesium, 9Calcium carbonate, 10Calcium, 11Total dissolved solids, 12Turbidity.

2.2. Data Analysis (LASSO Regression)
LASSO regression is a linear regression using shrinkage. The shrinkage is where the data value shrinks towards the center point (for example, the average value). The LASSO program encourages the use of simple, sparse models (models with fewer parameters). This special regression type is very useful for models that show high multicollinearity or when you want to automate certain parts of the model selection (such as variable selection/parameter elimination) [10]. The acronym “LASSO” stands for the least absolute shrinkage and selection operator. LASSO regression aims to obtain the prediction subset that minimizes the prediction error of the quantitative response variable. LASSO achieves this by imposing constraints on the model parameters, which will reduce the regression coefficient of a specific variable to zero [11]. Formally define the method of linear regression model:

\[ Y = X\beta + \varepsilon \]  

(1)
Where:

- $Y$: A vector represents the variable response observations of the class (N x 1);
- $X$: The matrix represents the observations of the explanatory variables of the type (N x p).
- $\beta$: Vector parameters are estimated from the class 
- $\mathbf{e}$: Random error vector of the class (N x 1) is considered by assuming $E(e) = 0$, $Var(e) = I\sigma^2$.

The columns of $X$ are denoted as $(X_1, X_2, ..., X_p)$ These represent the $P$ independent variables. The LASSO estimate of $\beta$

$$
\hat{\beta}^* = \min\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij}) + \lambda \sum_{j=1}^{p} |\beta_j| \}
$$

(2)

Where:

$$
\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij}) \quad \text{Residual sum of squares.}
$$

$$
\sum_{j=1}^{p} |\beta_j| \quad \text{LASSO penalty.}
$$

For estimating the parameter in the above equation, the R program (version 3.4.3) and LASSO package will use [12,13].

---

**Table 2. The quality measurement results of drinking water for imported items [4].**

| Symbol | Mn$^\text{a}$ | Fe$^\text{b}$ | Cl$^\text{c}$ | NO3$^\text{d}$ | F$^\text{e}$ | Na$^\text{e}$ | SO4$^\text{f}$ | pH |
|--------|--------------|--------------|-------------|-------------|---------|---------|----------|----|
| I1$^i$ | 0.0005       | 0.010        | 20          | 4.1         | 0.6     | 20      | 31       |    |
| I2$^i$ | 0.0000       | 0.000        | 22          | 5           | 0.9     | 22      | 25.6     |    |
| I3     | 0.0000       | 0.005        | 18          | 0.2         | 0.2     | 17      | 54       |    |
| I4     | 0.001        | 0.000        | 56          | 1.3         | 0.7     | 11      | 19       |    |
| I5$^i$ | 0.0000       | 0.015        | 21          | 0.6         | 0.7     | 19      | 33       |    |
| I6$^i$ | 0.0000       | 0.010        | 21          | 7.1         | 0.6     | 9       | 34       |    |
| I7$^i$ | 0.0000       | 0.010        | 15          | 4.51        | 0.8     | 12      | 51.5     |    |
| I8$^i$ | 0.0000       | 0.010        | 4.5         | 0.1         | 0.2     | 1       | 2.2      |    |
| I9$^i$ | 0.0000       | 0.000        | 43          | 6           | 0.67    | 19      | 5.95     |    |
| I10$^i$| 0.0000       | 0.010        | 22.15       | 6.9         | 0.85    | 21      | 14       |    |
| KSA Standard | 0.05 | 0.300 | 250 | 10 | 0.6-1 | - | 250 |   |

| Symbol | Mg$^\text{a}$ | CaCo$^\text{b}$ | Ca$^\text{c}$ | TDS$^\text{d}$ | TU$^\text{e}$ | pH |
|--------|--------------|-----------------|--------------|----------------|-------------|----|
| I1$^i$ | 5.05         | 46              | 10           | 118            | 0.28       | 7.2 |
| I2$^i$ | 7.7          | 45.5            | 5.5          | 105            | 0.23       | 7.2 |
| I3$^i$ | 19.5         | 105             | 10           | 120            | 0.18       | 7.8 |
| I4$^i$ | 10.55        | 101             | 23           | 111            | 0.1        | 7.65 |
| I5$^i$ | 1.8          | 50              | 22.5         | 112            | 0.22       | 7.4 |
| I6$^i$ | 6.45         | 121.5           | 38           | 159            | 0.26       | 7.7 |
| I7$^i$ | 13.4         | 95              | 16.5         | 114            | 0.115      | 7.2 |
| I8$^i$ | 5.55         | 50              | 10.7         | 50             | 0.23       | 7.65 |
| I9$^i$ | 17.6         | 110             | 15           | 139            | 0.4        | 7.3 |
| I10$^i$| 8.5          | 75              | 16           | 120            | 0.16       | 7.7 |
| KSA Standard | 30 | 300 | 75 | 100-700 | 5 | 6.5-8.5 |

$^a$Imported, $^b$the Kingdom of Saudi Arabia, $^c$Manganese, $^d$Iron, $^e$Chloride, $^f$Nitrate, $^g$Fluoride, $^h$Sodium, $^i$Sulphates, $^j$Magnesium, $^k$Calcium carbonate, $^l$Calcium, $^m$Total dissolved solids, $^n$Turbidity.
3. Results and discussions

Many researchers have been using LASSO regression as a significance test [14, 15,16,17,18,19]. Both local and imported bottled drinking water, pH element was selected as y parameter, and others were selected as x parameters. The value (Beta) for imported and local elements, as shown in table (3). In the case of imported items, the results showed that the elements (Nitrate, Sodium, Sulphates, and Magnesium) are insignificant (Beta > 0.05) while the elements (Iron, Chloride, Fluoride, Calcium carbonate, Calcium, Total dissolved solids, and Turbidity) are significant (Beta < 0.05). For the local items, the results showed that the elements (Nitrate, Sulphates, and Magnesium) are insignificant (Beta > 0.05) while the elements (Iron, Chloride, Fluoride, Sodium, Calcium carbonate, Calcium, Total dissolved solids, and Turbidity) are significant (Beta < 0.05).

Table 3. The Beta elements value of bottled drinking water for imported and local items.

| Elements | Named on Program | Beta (Imported Items) | Beta (Local Items) |
|----------|------------------|-----------------------|-------------------|
| Mn<sup>1</sup> | VAR001 | 0.000 | 0.000 |
| Fe<sup>2</sup> | VAR002 | 0.000 | 0.000 |
| Cl<sup>3</sup> | VAR003 | 0.000 | 0.000 |
| NO<sub>3</sub><sup>4</sup> | VAR004 | 0.300 | 0.268 |
| F<sup>5</sup> | VAR005 | 0.000 | 0.000 |
| Na<sup>6</sup> | VAR006 | 0.100 | 0.000 |
| SO<sub>4</sub><sup>7</sup> | VAR007 | 0.201 | 1.684 |
| Mg<sup>8</sup> | VAR008 | 0.416 | 0.062 |
| CaCO<sub>3</sub><sup>9</sup> | VAR009 | 0.000 | 0.000 |
| Ca<sup>10</sup> | VAR010 | 0.000 | 0.000 |
| TDS<sup>11</sup> | VAR011 | 0.000 | 0.000 |
| TU<sup>12</sup> | VAR012 | 0.000 | 0.000 |

<sup>1</sup>Manganese, <sup>2</sup>Iron, <sup>3</sup>Chloride, <sup>4</sup>Nitrate, <sup>5</sup>Fluoride, <sup>6</sup>Sodium, <sup>7</sup>Sulphates, <sup>8</sup>Magnesium, <sup>9</sup>Calcium carbonate, <sup>10</sup>Calcium, <sup>11</sup>Total dissolved solids, <sup>12</sup>Turbidity.

4. Conclusions

This paper has presented a methodology by analyzing LASSO regression for both imported and local bottled drinking water. From the results which have been obtained, we observed that LASSO regression gave better results. LASSO regression is to get the subset of predictors that minimizes prediction error for a quantitative response variable.

5. References

[1] Bhawana J, Singh A and Susan M 2019. The world around bottled water. *Bot. and Pack. Wat.*, 4, pp. 39-61.
[2] Huerta-Saenz L, Irigoyen M, Benavides J and Mendoza M 2012. Tap or bottled water: drinking preferences among urban minority children and adolescents. *Jour. of Comm. Heal.*, 37(1), pp.54-58.
[3] Johnstone N and Serret Y 2012. Determinants of bottled and purified water consumption: results based on an OECD survey. *Wat. Pol.*, 14(4), pp. 668-679.
[4] Hussein A and Mohammed R 2019. The Quality of Drinking Water Bottled Domestic and Imported in Iraq. *Jour. of Eng. and App. Sci.*, 14(9), pp.10572-10578.
[5] Tibshirani R 1996. Regression shrinkage and selection via the lasso. *Jour. of the Ro. Stat. So: Series B (Methodological)*, 58(1), pp.267-288.
[6] Tibshirani R 2011. Regression shrinkage and selection via the lasso: a retrospective. *Jour. of the Roy. Stat. Soc.: Series B (Statistical Methodology)*, 73(3), pp.273-282.
[7] Hussein A, 2017. Azo textile dyes wastewater treatment with constructed wetlands: design and operation of experimental vertical-flow constructed wetlands applied for the treatment of azo textile dyes (with/without artificial wastewater) (Doctoral dissertation, University of Salford).
[8] Hussein A and Scholz M 2017. Dye wastewater treatment by vertical-flow constructed wetlands. *Eco. Eng.*, **101**, pp.28-38.

[9] Hussein A and Scholz M 2018. Treatment of artificial wastewater containing two azo textile dyes by vertical-flow constructed wetlands. *Env. Sci. and Poll. Res.* pp. 1-20.

[10] Chakraborty A, Bhattacharya A, and Mallick BK 2020. Bayesian sparse multiple regression for simultaneous rank reduction and variable selection. *Biom.*, **107**(1), pp.205-221.

[11] Vidyasagar M 2015. Identifying predictive features in drug response using machine learning: opportunities and challenges. *Annual review of pharmacology and toxicology*, **55**, pp.15-34.

[12] Yi C and Huang J 2017. Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. *Jour. of Comp. and Graph. Stat.*, **26**(3), pp. 547-557.

[13] Reid S, Tibshirani R, and Friedman J 2016. A study of error variance estimation in lasso regression. *Stat. Sin.*, pp.35-67.

[14] Lockhart R, Taylor J, Tibshirani R J and Tibshirani R, 2014. A significance test for the lasso. *Ann. of stat.*, **42**(2), p.413.

[15] Buja A, and Brown L 2014. Discussion: A significance test for the lasso. *The Ann. of Stat.*, **42**(2), pp.509-517.

[16] Tabassum M and Ollila E 2017. Pathwise least angle regression and a significance test for the elastic net. 25th *Euro. Sig. Proc. Conf. (EUSIPCO)*, pp. 1309-1313.

[17] Li F, Yang Y and Xing E P 2006. From lasso regression to feature vector machine. *In Adv. in Neu. Infor. Proc. Sys.*, pp. 779-786.

[18] Fonti Vand Belitser E 2017. Feature selection using lasso.*VU Amst. Res. Paper in Bus. Anal.*, **30**, pp.1-25.

[19] Hastie T, Taylor J, Tibshirani R and Walther G 2007. Forward stagewise regression and the monotone lasso. *Elec. Jour. of Stat.*, **1**, pp.1-29.