Filip Carević

Report for Applied Biostatistics Course MATH-493

Developing coal pillar stability chart using logistic regression

1. Introduction

In the underground rock and coal mining industries, pillars play a significant structural role. They conduct tunneling and mining work with temporary or permanent support. Rock cracks can lead to roof collapses and rock bursts if the pillars are unstable. In addition, increase of the mining depth increases more frequent pillar instabilities as well, leading to more common accidents. Taking into account that life of the mine lasts several years, or even decades, we may conclude that the pillar stability is a crucial issue that must be addressed.

In this paper, we follow more recent approaches which aim to statistically examine the pillar instability problem. We utilize logistic regression model to predict the certainty of pillar being classified as stable.

2. Exploratory data analysis

2.1. Data description

The data used for the logistic regression has been collected by [1] and represents historical cases of the coal pillars. In addition, the data has been separated into two groups- *stable* (indicated as 1) and *failed* (indicated as 0).

Each datapoint is described with the following features:

- **Pillar ID** - Unique identifier of the pillar.
- **Mine Seam** - The mine from where the pillar case has been recorded.
- **Depth** - Depth of the mine, represented in meters [m].
- **Height** - Height of the pillar, represented in meters [m].
- **Width** - Width of the pillar, represented in meters [m].
- **Width to Height ratio** - Feature obtained from the division of *width* and *height* features.
- **Roadway width** - The width of the roadway inside the observed mine, represented in meters [m].
- **Uniaxial compression strength** - Uniaxial compression strength of coal, represented in Megapascal [MPa].
- **Strength** - Strength of the pillar, represented in Megapascal [MPa].
- **Stress** - Stress of the pillar, represented in Megapascal [MPa].

- **Strength to Stress ratio** - Feature obtained from the division of *strength* and *stress* features.
- **Stability** - Indicator of stability, where stable pillars are indicated as 1, while failures are indicates as 0.

2.2. Data analysis

Before performing data analysis, the irrelevant features are removed. Features that are considered irrelevant are *Pillar id* and *Mine Seam*, since they do not contain any information significant for the outcome prediction.

In order to gain a better insight on the data, we examine conditional density plots of the outcome (i.e, *Stability*) and corresponding feature. The plots have been presented on the figure 2.1. Interesting tendency can be observed on the relationship between *Stability* and *Height to Width ratio*, where increase in portions of stable pillars follows the increase in observed feature. Similar trends can be noticed on the plots of outcome and *Strength to Stress ration*, as well as outcome and *Strength*.On the other hand, the relationship between outcome and *height* depicts that with increase of the height, the portion of stable pillars decreases. In addition, extraordinary behaviour could be noticed in relations of outcome with *width*, *depth* and *stress*, respectfully, where fluctuations in the pillar stability may be explained with outliers in the dataset. However, since the dataset size is relatively small (only 29 datapoints), we are not able to make any conclusion for the cause of this irregularities. Other plots do not express any intriguing patterns.

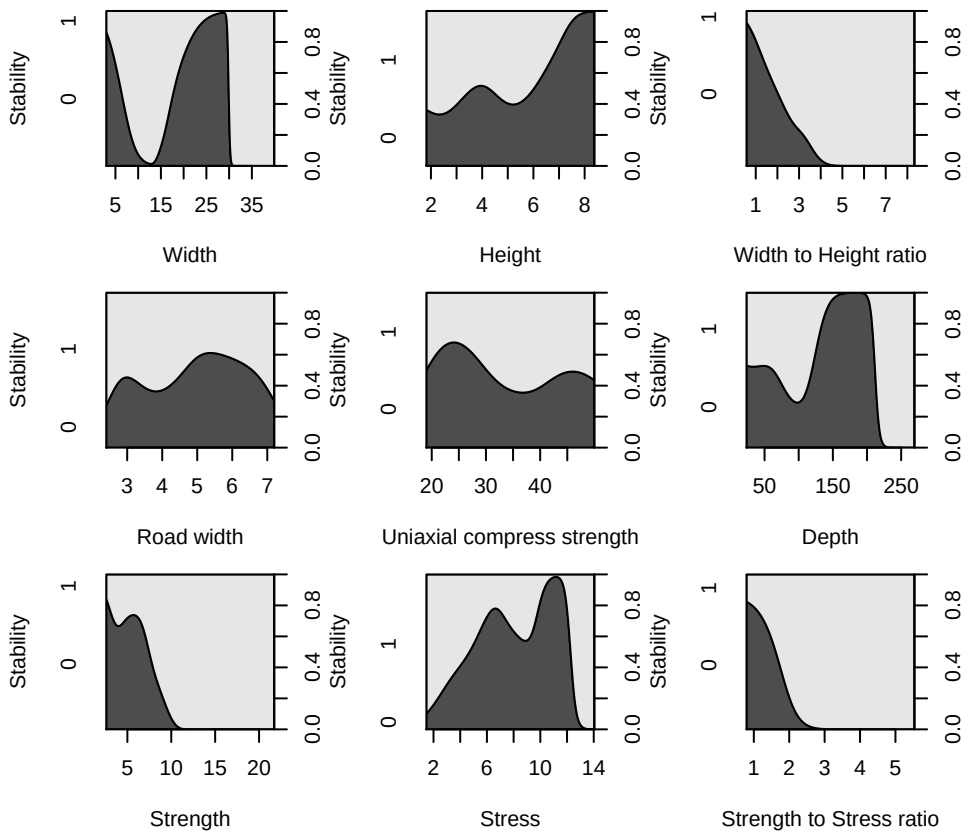

Figure 2.1: Graphical bivariate exploratory data analysis

3. Model assessment - logistic regression

3.1. Definition of the logistic regression

Logistic regression is a statistical model used to model the probability of a categorical outcome given one or more input variables (i.e, features). It is used for the binary classification problems, i.e, problems where outcome could be encoded with an indicator values such as: true/false, yes/no, one/zero and similar. However, note that it does not predict a label (i.e, does not perform a classification). It predicts a probability associated with a corresponding label which is a real value, hence the name logistic *regression*. It can be generalized for multi-class classification problems, where it is called *multinomial logistic regression*.

Binary logistic regression model estimates the logarithm of the odds (i.e, log-odds) of the event happening as an linear combination of the input features. By definition, for the input with n features, the logistic regression is defined as:

$$Pr(Y_i = 1|X_i) = \frac{exp(\beta_0 + \beta_1 X_i + \beta_2 X_2 + ... + \beta_n X_n)}{1 + exp(\beta_0 + \beta_1 X_i + \beta_2 X_2 + ... + \beta_n X_n)}$$

Parameters of the model are estimated using the approach called the *maximum likelihood estimation*(MLE). This approach aims to estimate the model weights $\beta = (\beta_0, \beta_1, ..., \beta_n)$, which maximize the probability obtained on the observed data. Under the assumption that the datapoints are independent, the likelihood function of the data given the model weights is defined as:

$$\mathrm{l}(\beta) = p(\mathbf{y}|\mathbf{X}, \beta) = \prod_{n=1}^{N} p(y_n|x_n, \beta) = \prod_{n:y_n=1} p(y_n = 1|x_n, \beta) \prod_{n:y_n=0} p(y_n = 0|x_n, \beta)$$

$$l(\beta) = p(\mathbf{y}|\mathbf{X}, \beta) = \prod_{n=1}^{N} \sigma(x_n^T \beta)^{y_n} \sigma(x_n^T \beta)^{1-y_n}$$

To simplify the equation to a simpler, the logarithm is applied, which gives us the form:

$$L(\beta) = ln(l(\beta)) = \sum_{n=1}^{N} y_n ln(\sigma(x_n^T \beta)) + (1 - y_n) ln(1 - \sigma(x_n^T \beta))$$

In order to find the optimal values of weights β which maximize the log-likelihood function $\mathrm{L}(\beta)$, previous equation is differentiated with respect to weights β. Later, to find the stationary point, the resulting equations, the first one representing the derivative with respect to intercept and the second one represents the derivative with respect to weight β, are set to zero:

$$\sum_{n=1}^{N} y_n - \sigma(x_n^T \beta) = 0 \quad \sum_{n=1}^{N} x_{nj}[y_n - \sigma(x_n^T \beta)] = 0$$

Note that, unlike the plain linear regression, the aforementioned equations do not have the closed-form solution. Instead, they have to be solved using iterative methods, such as *Gradient Descent* or *Newton method*.

3.2. Model selection

For the model selection procedure, forward feature selection approach is selected. The Akaike information criterion (AIC) estimator was used to determine the best models in observed iteration.

The procedure converged after the two iteration (i.e., selection of two features), since obtained model was able to perfectly separate data. **Strength to Stress ratio** (*ss_ratio*) and **Strength** (*strength*) features, including the intercept, were preserved, and the final model is defined as:

$$p(y = 1|ss\_ratio, strength) = \frac{1}{1 + exp(-[-3944.75 + 1481.43 * ss\_ratio + 276.76 * strength])}$$

Note that the coefficient values are large, which is expected, since the data is linearly separable. In case where smaller model weights are preferred, the regularization should be added. After the addition of Lasso regularization, we obtain model depicted in table 3.1.

Table 3.1: Lasso logistic regression

Feature	Coefficient
Intercept	0.02
Strength to Stress ratio	4.13
Strength	1.29

3.3. Model assessment

After the model selection, we need to check if important assumptions of logistic regression are satisfied:

1. **Binary outcome**

   Since our outcome is *pillar stable* or *pillar failure*, we conclude that this supposition is satisfied.

2. **Absence of multicollinearity**

   In order to detect multicolinearity in model features, the utilize variation inflation factor (VIF). Values of VIF for selected model are present in the table 3.2. Since no values are higher than commonly used threshold of 5, we may conclude that no multicollinearity is present. In other word, this supposition is satisfied.

Table 3.2: Variation inflation factor (VIF)

Feature	VIF
Strength to Stress ratio	3.51
Strength	3.51

3. **Absence of outliers**

   The dataset consists only of 29 samples. Therefore, it is hard to define what and outlier really is. We could be certain that a datapoint is outlier only if its feature values are extreme (e.g., several orders of magnitude higher than expected). However, these special cases do not exist in our dataset.

4. **Independent observations**

   Since datapoints represent individual pillars, and no two datapoints are derived from the same pillar, we may conclude that this assumption holds.

5. **Linear relation between logit and linear predictors**

   This supposition is satisfied, as can be seen on the plotted relationships:

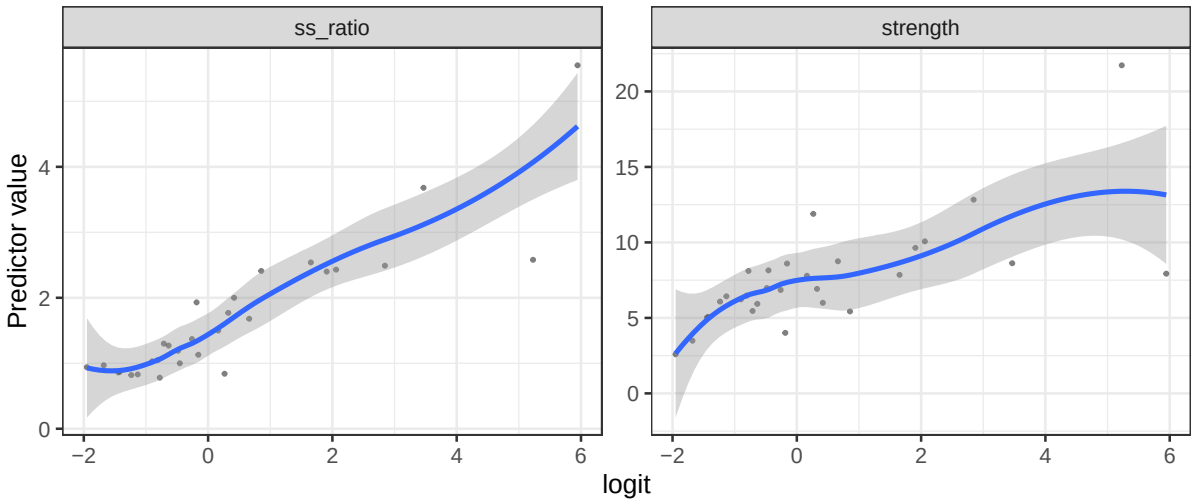

Figure 3.1: Relationship between logit of the outcome and predictor variables

4. Conclusion

In this paper, we utilized the logistic regression in order to assess the stability of pillars inside the mines. We showed that the best model utilizes the information contained in **Strength to Stress ratio** and **Stress** features. Moreover, the model manages to completely separate the data. Lastly, we proved that crucial logistic regression assumptions are satisfied for the aforementioned model.

Bibliography

[1] R.K. Wattimena, S. Kramadibrata, I.D. Sidi, and M.A. Azizi. Developing coal pillar stability chart using logistic regression. *International Journal of Rock Mechanics and Mining Sciences*, 58:55–60, 2013.