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 [7] 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 features *Pillar ID* and *Mine Seam*, since they do not contain any information significant for the outcome prediction, are removed.

Table 2.1: Univariate numerical analysis

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Depth	23.00	36.00	54.00	77.79	90.00	270.00
Height	1.80	3.00	3.60	4.16	5.10	8.40
Width	2.90	4.70	6.30	8.99	9.90	40.00
W/H ratio	0.57	1.41	1.8	2.27	2.86	8.33
Roadw. width	2.40	4.20	5.40	5.07	6.00	7.20
Uniax. compress.	19.00	26.00	33.00	34.31	43.00	50.00
Strength	2.59	5.93	6.92	7.62	8.60	21.73
Stress	1.43	3.61	5.15	5.37	6.17	14.08
S/S ratio	0.78	0.97	1.30	1.70	2.40	5.55

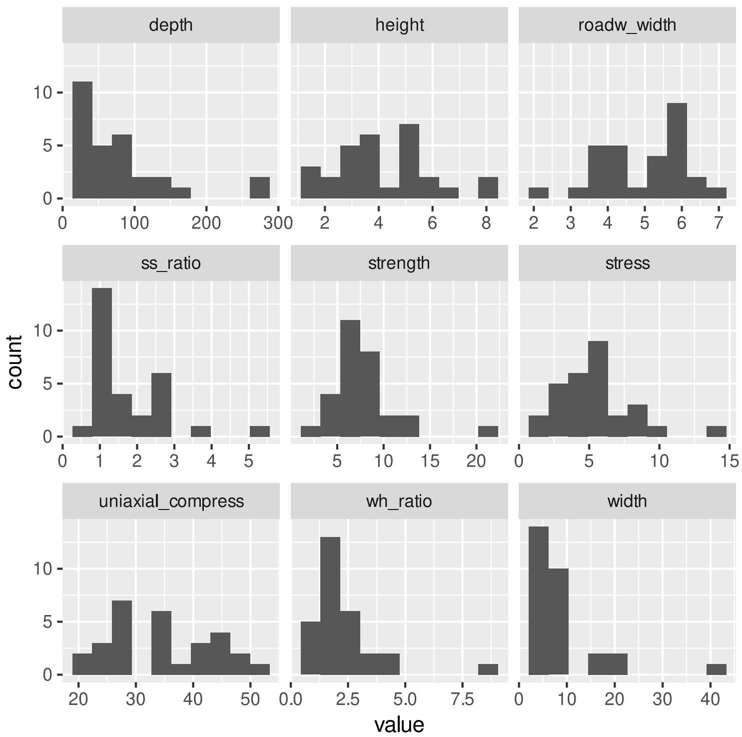

Figure 2.1: Histogram plots

Table 2.1 and figure 2.1 depict common univariate statistics for each feature. It is intriguing to note that *Depth* and *Width* features are characterized by a noticeably greater mean than median. Moreover, from corresponding histograms (figure 2.1), we may see the possible existence of outliers. We will check for existence of outliers later in the section 3.3.

Interesting tendencies can be observed on the conditional density plots (figure 2.2) of the outcome (i.e, *Stability*) and corresponding features. The relationship between *Stability* and *Height to Width ratio* shows us that 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 ratio*, 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. It is worth to note extraordinary behaviour noticed in relations of outcome with *Width* and *Depth*, respectfully, where fluctuations in the pillar stability could be explained with the existence of outliers in the dataset. This may support the hypothesis stated in the previous paragraph.

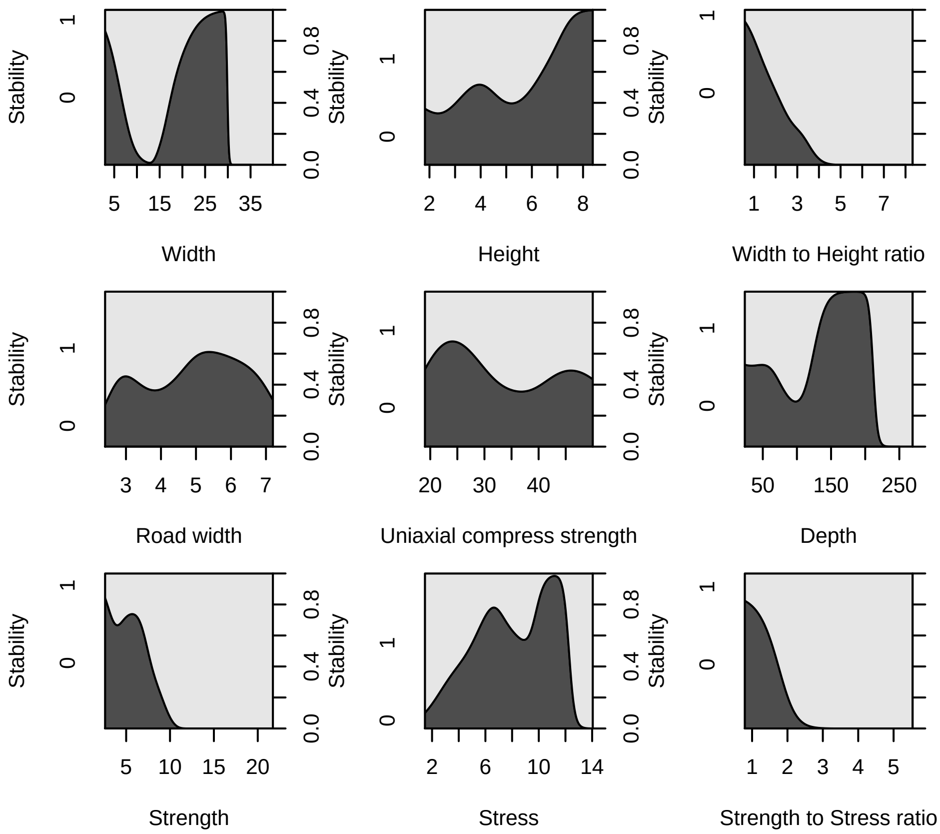

Figure 2.2: Graphical bivariate exploratory data analysis

The figure 2.3 portrays the data separated by *Stability* (i.e., outcome) feature - the blue color represents class 1, while the red color represents class 0. The correlation subplots present Pearson's correlation coefficients [2] for corresponding covariates, where we can observe several highly correlated features, such as *Depth* and *Stress*, *Depth* and *Strength*, etc. The histogram plots support the observations made in the previous paragraph. However, the most interesting behaviour is presented on the scatter plots, where the tendency for data separation in two different clusters is exhibited. This tendency is more visible in the *WH_ratio-SS_ratio*, *Stress-SS_ratio* and *Strength-SS_ratio* plots. Moreover, the *Strength-SS_ratio* plot denotes us that the data may be linearly separable.

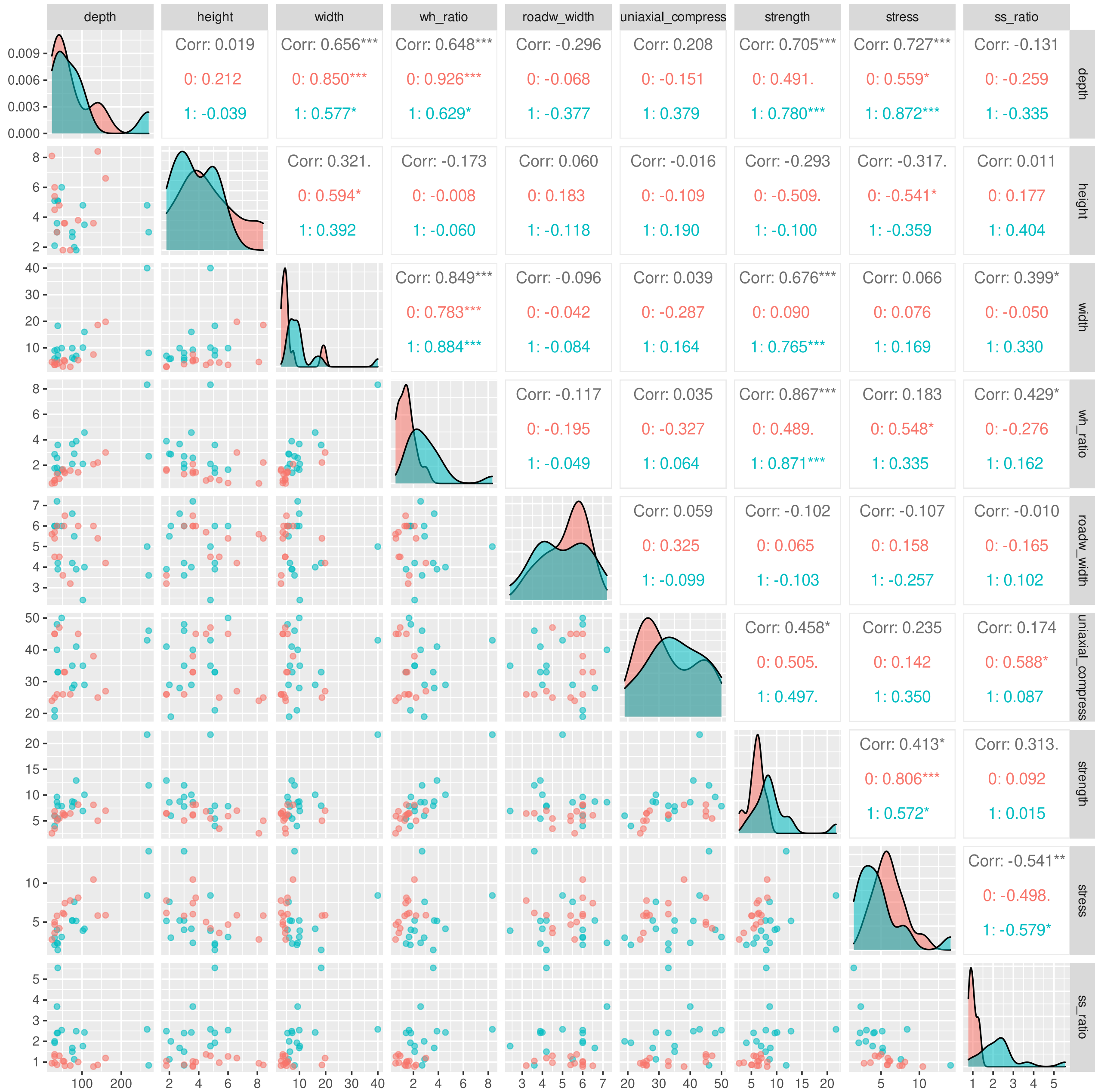

Figure 2.3: Scatter, histogram and correlation plots

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 a 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:

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

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

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 $L(\beta)$, 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, the forward feature selection approach is used. The Akaike information criterion (AIC) [6] estimator was utilized to determine the best models in observed iteration. The Akaike information criterion depicts the amount of information lost by the observed model. The less information is lost (i.e., the lower AIC), the model better explains the data. Formally, it is defined as:

$$AIC = 2k - 2ln(\hat{L})$$

where k represents the number of estimated parameters, and $\hat{L}$ represents the maximum value of the likelihood function for the observed model.

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 (described in [4]) 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 multicollinearity in model features, we utilize variation inflation factor (VIF). Values of VIF for selected model are present in table 3.2. Since no values are higher than commonly used threshold of 5, we may conclude that no multicollinearity is present. In other words, 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**

In order to investigate the existence of outliers in our dataset, firstly, we will inspect the existence of influential values. Influential values are extreme datapoints which could significantly affect the quality of obtained logistic regression model. To assess whether the individual datapoint is influential or not, we will use Cook's distance [5] metric. Cook's distance summarizes how much the model changes when the observed datapoint is removed from the dataset. Formally it is defined as:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE}$$

where

- $\hat{Y}_j$ is the prediction for the j^{th} observation
- $\hat{Y}_{j(i)}$ is prediction for the j^{th} observation using the model trained on dataset after excluding i^{th} observation
- p denotes total number of model parameters
- MSE denotes model's Mean Squared Error[1]

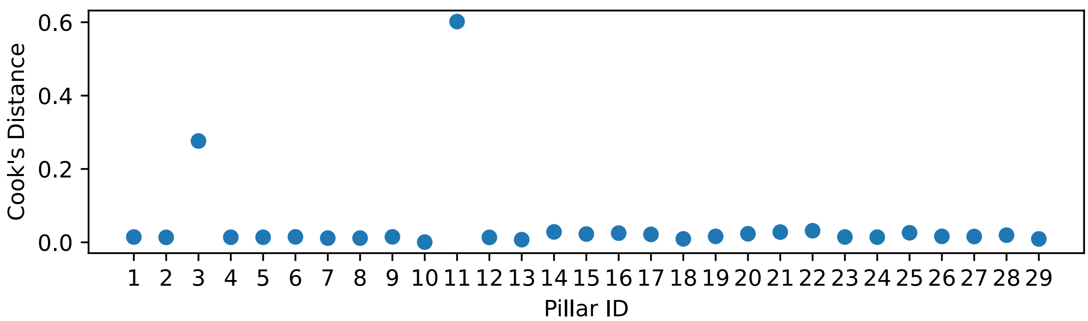

Figure 3.1: Cook's distance plot

As a rule of thumb, the datapoint is considered to be influential if the distance is greater than $4/(N + p - 1)$, where N denotes number of samples, and p denotes number of estimated model parameters. In our case, the threshold is 0.13. From the figure 3.1, we can see that datapoints represented with IDs 3 and 11 have influence greater than the calculated threshold.

To be certain that those influential values are outliers, we will assess the standardized residuals plot. The residual represents the difference between the observed and predicted value, whereas standardized residual is the residual obtained after division with standard deviation calculated over all residuals [3]. The plot portrays standardized residuals of corresponding data samples. Note that the data sample is represented with its *Pillar ID*.

An observation with standardized residual with absolute value greater than 3 (the threshold for 95% confidence interval for the Student's t-distribution with corresponding degrees of freedom) is considered to be the outlier [3]. The figure 3.2 shows absence of any such case, therefore we may conclude absence of outliers in our dataset.

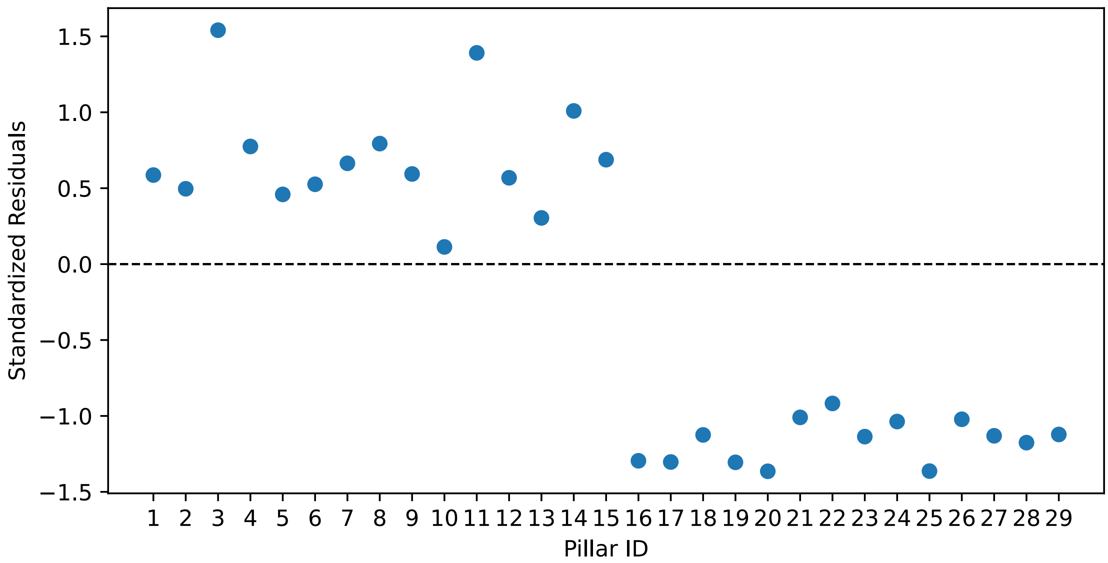

Figure 3.2: Standardized residuals plot

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:

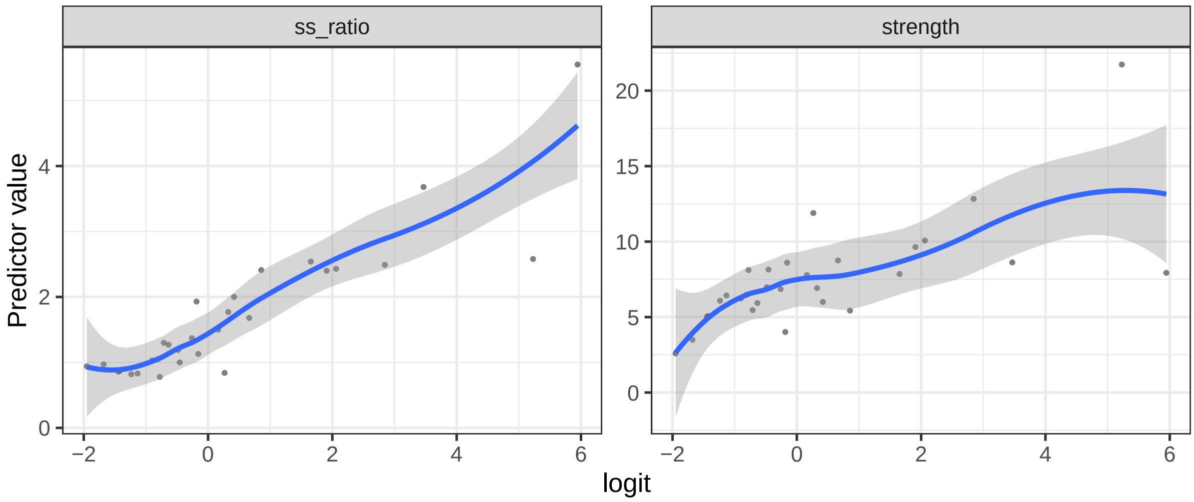

Figure 3.3: Relationship between logit of the outcome and predictor variables

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] David M. Allen. Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3):469–475, 1971.

[2] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[3] R. Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

[4] G David Garson. Testing statistical assumptions. *Asheboro, NC: Statistical Associates Publishing*, 2012.

[5] Choongrak Kim and Barry E. storer. Reference values for cook's distance. *Communications in Statistics - Simulation and Computation*, 25(3):691–708, 1996.

[6] Eric-Jan Wagenmakers and Simon Farrell. Aic model selection using akaike weights. *Psychonomic bulletin review*, 11:192–6, 03 2004.

[7] 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.