Logistic Regression Analysis in Gold targetting

C. Tissier

1 Introduction

From as early as 1500 BC, gold has been perceived as a valuable good and accepted as a satisfactory form of payment. History has given gold a power surpassing that of any other commodity on the planet, and that power has never really disappeared. Gold is both a metal and a mineral extracted from deep under the surface. Today, with the development of statistical theory and bigdata it is natural to optimize gold targetting using geological data.

Research questions and approaches

Do water/chemical factors as As, Sb influence the presence of gold deposit ? Does the probability of having a gold deposit augment in proximity of a lineament?

In this report, we will do an exploratory data analysis and a Logistic regression analysis in order to give an answer to the research questions.

2 Dataset

This paper is inspired by the work of Nihar Ranjan Sahoo and Hari Shankar Pandalai Sahoo and Pandalai (1999). The dataset used in this report can be found online [1]. It consists of 64 observations of four variables:

- `As` is a continuous variable representing the level in ppm (part per million) of the chemical element Arsenic also refered as As in the periodic table.
- `Sb` is a continuous variable representing the level in ppm (part per million) of the chemical element Antimony also refered as Sb in the periodic table.
- `Lineament_proximity`: a lineament is a linear feature in a landscape which is an expression of an underlying geological structure. Here Lineament_proximity is a binary variable that takes value 1 if lineament can be found in a radius of 5 km and 0 otherwise.
- `Gold_proximity` is our dependant variable. It is binary and takes value 1 if gold deposit can be found in a radius of 5 km and 0 otherwise.

[1] `https://users.stat.ufl.edu/~winner/data/gold_target1.dat`

3 Exploratory data analysis

3.1 Univariate EDA

In total, our data contains 36 observations with variable $\widehat{Gold\_proximity}$ equal to 0 and 28 observations with variable $\widehat{Gold\_proximity}$ equal to 1 (see Figure1). Moreover, there is a balanced distribution for the variable Lineament_proximity with 32 observations for both level of this factor.

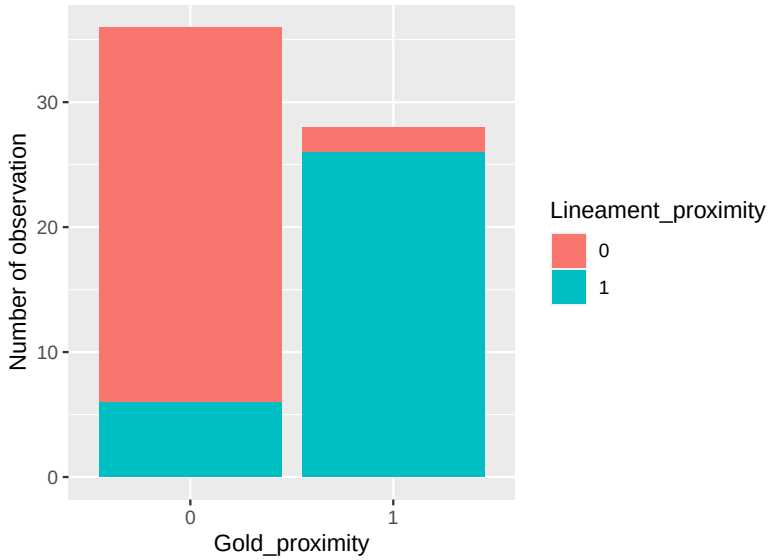

Figure 1: Histograms of the number of observations depending on the presence of gold deposit, with the the Lineament_proximity proportion

The As variable takes value ranging from 0.1ppm to 41.480ppm with a mean of 4.645 and a median of 1.235 ppm.

The Sb variable takes value ranging from 0.1ppm to 18.200ppm with a mean of 2.039 and a median of 0.650 ppm.

3.2 Bivariate EDA

We can observe from Figure2 that both As and Sb level seem to be concentrated around 0 when there is absence of a near gold deposit whereas the first quantile in presence of a near gold deposit is around 5ppm for the level of As and is around 3ppm for Sb. This suggests a positive correlation between the level of both As and Sb and the presence of a gold deposit.

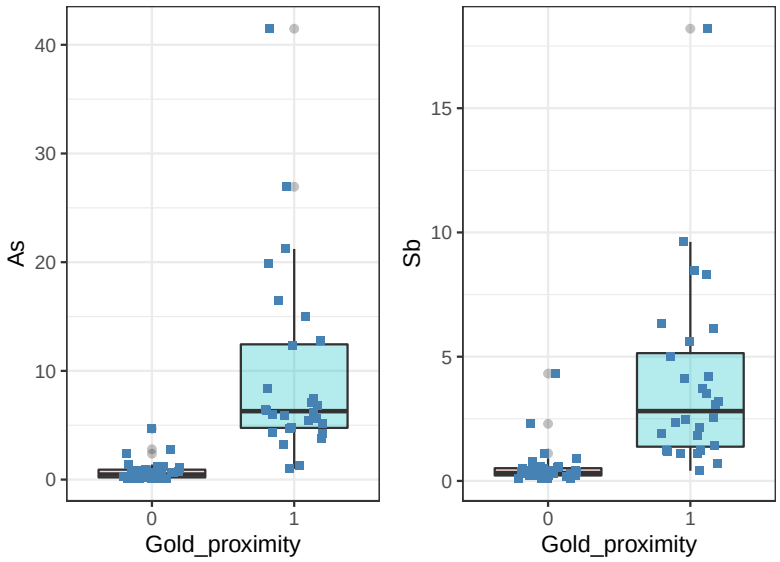

Figure 2: Boxplot of the As and Sb levels with respect to proximity to a gold deposit.

Also, the correlation between the As and Sb variables is 0.7459591.

4 Logistic regression analysis

4.1 Logistic regression fitting

The model was chosen by forward selection trying to get the best AIC. From the null model we first added the variable As which gives us an AIC of 26.60. Then adding the Sb variable we got an AIC of 25.95 and lastly we reached an AIC of 22.294 by adding the Lineament_proximity variable. Adding interaction terms didn't improve the AIC so our final model is:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 As + \beta_2 Sb + \beta_3 Lineament\_proximity + \epsilon \tag{1}$$

where $\pi(x)$ is the probability of the dependant variable $\widehat{Gold\_proximity}$ to be equal to 1 assuming predictors x where x refers to As, Sb and $Lineament\_proximity$. ϵ is the error term.
The coefficient estimates can be found in Table 1

We can note that the p-value for both β_2 and β_3 are lesser than 0.05. Nonetheless we kept both the Sb and the Lineament_proximity variables as by doing a likelihood ratio test with a model that doesn't contain them we see that they are statistically significant at level 0.05 and our final model is:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = -7.61 + 1.20 As + 1.42 Sb + 3.20 Lineament\_proximity + \epsilon \tag{2}$$

Parameter number	**Parameter estimate**	**Parameter standard error**	z **value**	$\mathbf{P}(>\|z\|)$
β_0	-7.61	3.17	-2.40	0.0162
β_1	1.20	0.49	2.46	0.0139
β_3	1.42	0.73	1.95	0.0516
β_4	3.20	1.89	1.69	0.0909
Summary statistics				
Number of observations = 64				
$\mathcal{L}(0) = -43.86$				
$\mathcal{L}(\hat{\beta}) = -7.10$				
AIC = 22.19				

Table 1: Summary table for the logistic regression

4.2 Logistic regression interpretation

To interpret our result, we first need to compute the Odd Ratio (OR), that is we compute the exponential of our estimated coefficients. We give them in Table 2.

We can first note that the confidence intervals are pretty wide. This is because we only have 64 observations. The odd ratio for the intercept is 4.95×10^{-4}, this is the odd of having a near gold deposit for an observation with both As and Sb level of 0 and no Lineament proximity.

The odd ratio for the continuous variables (As and Sb) are larger than one which indicates that the odd of a near gold deposit augments when both the level of As and Sb augments.

Lastly the odd ratio of the binary variable Lineament_proximity is also greater than one which indicates that the odd of a near gold deposit is higher for observations with $Lineament\_proximity = 1$

Parameter number	**Parameter estimate**	**OR**	**OR CI lower 2.5%**	**OR CI upper 97.25 %**
β_0	-7.61	4.95×10^{-4}	2.96×10^{-8}	2.59×10^{-2}
β_1	1.20	3.34	1.64	14.96
β_3	1.42	4.14	1.28	32.28
β_4	3.20	2.45	1.10	3963.55

Table 2: Odd ratio for the logistic regression model

4.3 Model assessment

In order to draw some conclusions from our logistic regression model, certain assumptions must be checked. In particular, we would like to verify the linearity of independent variables and log-odds. Therefore we will inspect the six following assumptions using diagnostic plots.

1. Binary outcome
   The dependent variable $\widehat{Gold\_proximity}$ can only take the value 0 and 1 so this assumption is verified.
2. Independent observations
   From the original paper Sahoo and Pandalai (1999) it is unclear how the data where gathered so it is hard to inspect this hypothesis. Nevertheless, We can note that there are no repetitions in the data and *a priori* the observations are indeed independents.
3. Linear relation between logit and linear predictor
   We first note that adding interaction terms led to worse model judging from the AIC, which can indicate that the relation between the log odd and the predictor is linear. Moreover we see in Figure 3 that the linearity assumption clearly hold for As. For Sb it is less clear but there are no evidence that the assumption is violated.

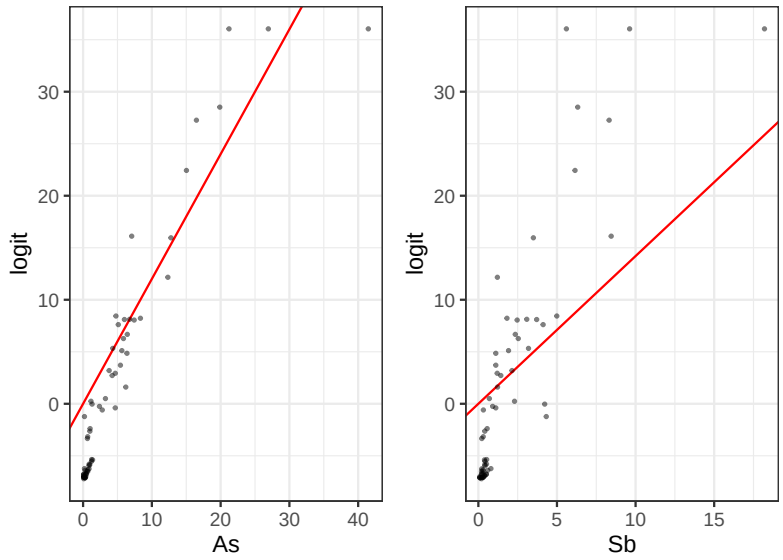

Figure 3: Scatter plot of logit vs the continuous predictors, with the line $y = \beta_1 \times As$ (resp.$y = \beta_2 \times Sb$) in red

4. No multicollinearity We calculated the VIF for each variables.
   - As: VIF=1.8
   - Sb: VIF=2.29
   - Lineament_proximity: VIF=1.87

   Since all are lesser than 5, the assumption of no multicollinearity is verified.
5. No outliers
   According to Figure 4 there are clearly some outliers so the assumption is not verified.

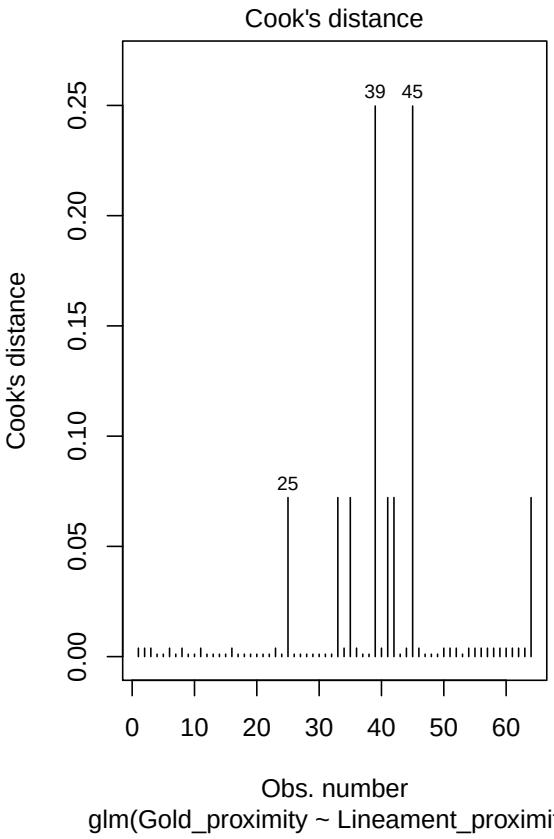

Figure 4: Cook distance for the logistic regression model

6. Large sample size
   This assumption is cleary violated as we only have 64 observations for three predictors.

5 Conclusion

In conclusion, we could observe that high level of Arsenic and/or Sb increase the chance of having the presence of a near gold deposit. The presence of a lineament in a radius of 5km also increase the chance of having the presence of a gold deposit in a radius of 5 km. Although all assumptions for a logistic regression were not satisfied (see Subsection 4.3), our model has a pretty good fit with a loglikelihood of -7.10 and an AIC of 22.19 (see Subsection 4.1).

6 References

Nihar Ranjan Sahoo and Hari Shankar Pandalai. Integration of sparse geologic information in gold targeting using logistic regression analysis in the hutti–maski schist belt, raichur, karnataka, india—a case study. *Natural Resources Research*, 8(3):233–250, 1999.