Leveraging Logistic Regression on Geologic Information for Gold Targeting

Applied Biostatistics

Kamran Nejad-Sattary

26 May 2021
Ecole Polytechnique de Lausanne

Introduction

The practice of exploration targeting, referring to estimating potential of mineral deposits in a defined geological setting, is at heart a Geological investigation founded on statistical analyses. Fundamentally, this remains a task of maximising mineral discovery potential while minimizing exploration costs. Coupled with the prospect of expensive minerals, such as Gold, mineral targeting appears an enticing geological practice to refine.

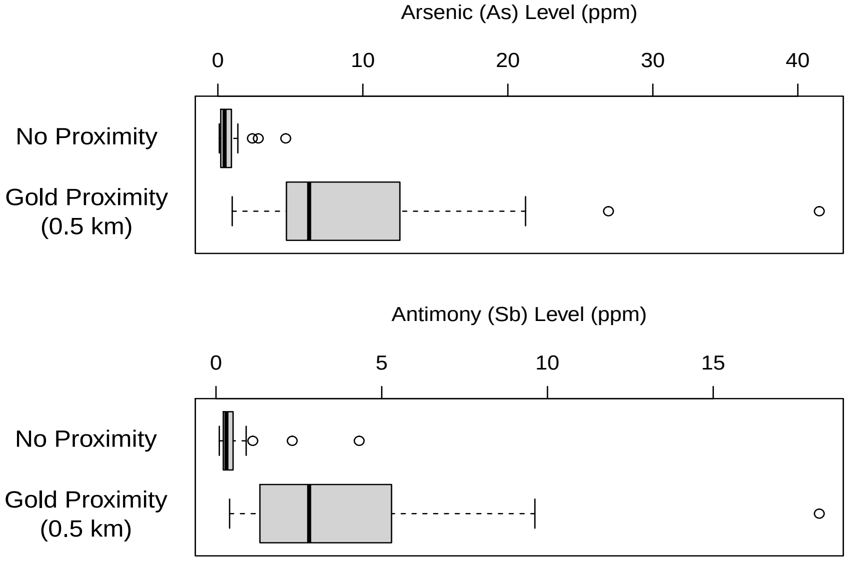
Owing to the previous work conducted by Sahoo and Pandalai, 1999, this study builds upon this original analysis which utilizes geochemical dispersion halos of Arsenic (As), Antimony (Sb), Mercury (Hg), and Bismuth (Bi) in soil, as well as As and Sb in groundwater, to statistically model the probabilities of gold deposits. This traditional practice has led to a divergence in employed techniques, ranging from data-driven and knowledge-driven approaches (Bonham-Carter, 1994). While an enormous variety of data-driven approaches have been tested (Sahoo & Pandalai, 1999, p. 234), this study explores the efficacy of the Logistic Regression in tackling this task, leveraging its ability to predict binary target variables using continuous and discrete data. For this, the As and Sb Levels, aswell as the proximity of Lineaments and Gold, recorded by Sahoo and Pandalai, 1999 from 80 analyzed soil samples collected just above the C horizon, are utilized.

Exploratory Data Analysis

Preliminary inspection into the utilized dataset serves a crucial first step into examining the employed explanatory variables. Figures 1 & 2 fulfill this purpose. It is observable that across the three chosen explanatory variables, there is minimal overlap in their distributions when comparing gold deposit to non-deposit locations. Specifically, for Arsenic and Antimony, the median value of the locations without gold do not lie within the whiskers of those with gold deposits, indicating that they lie outside the lower and upper quartiles of the gold-deposit locations, while for those locations close to gold deposits, the proximity of Lineament (which represents 32 or exactly half of the sample observations) seems to fairly reliably accompany the presence of Gold. Arsenic and Antimony levels are highly correlated, with a Pearson correlation of 74.60%, and the Arsenic and Lineament

Proximity Indicator are also highly correlated with a point-biserial correlation of 52.74%, although multicollinearity cannot yet be concluded. In fact, the variance inflation factor (VIF) which measures the ratio of a full model to that of a model explained by a single variable, employed on the later specified model, yields VIF scores of 1.58 and 2.29 respectively, not breaching the traditional 5 or 10 VIF threshold for concluding multicollinearity. It remains crucial to distinguish that the obtained sample ($N = 64$) remains small and potential variance from the sampling process may lead to presumptuous conclusions.

Figure 1: Distributions of Arsenic (As) and Antimony (Sb) Levels

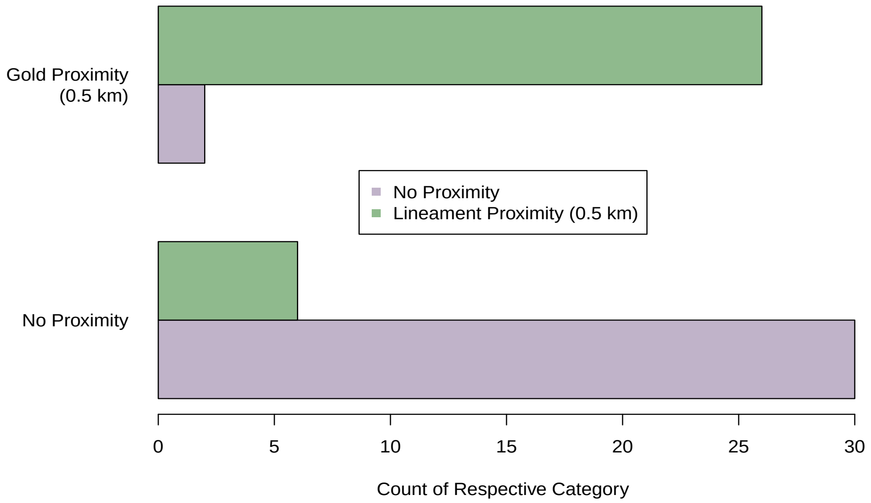


Further, it is noticeable that when Arsenic and Antimony levels are grouped by whether they are near gold, their values appear restricted to a smaller range of values. In fact, while the mean Arsenic level of the overall sample is 4.64, with values ranging from to 0.10 to 41.48, grouping by gold, which represents 28 out of the 64 data points, portrays a wildly different distribution of values. Specifically, those locations which were close to gold presented a mean Arsenic level of 9.65 with ranges of values between 0.99 to 41.48 in contrast to those without gold which accompanied a mean of 0.75, and a range of values from 0.10 to 4.68. This numerical uni-variate analysis is also observable in figure 1, which also reiterate this phenomenon for Antimony levels, on top of highlighting a positive skewness for each constructed group.

In addition, it is remarkable that the there exists so few outliers in the locations without gold which do have Arsenic and Antimony levels in the expected range of those locations with gold, specifically the locations without gold reliably have Arsenic and Antimony levels within a small distinct range than those with gold. These three explanatory variables seem to have strong explanatory power in our sample, though generalisability to out-of-sample data may be another matter.

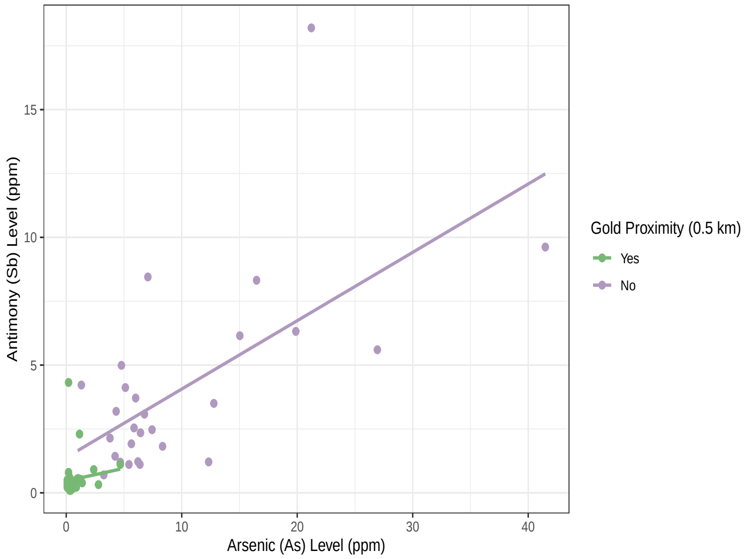
While figure 3 once again reemphasizes this near total separation between distributions of Arsenic and Antimony levels which are close and distant to gold, it also portrays the relationship between Arsenic and Antimony in the geochemical sampled soil locations. The points are dispersed though on average reveal an increasing relationship between the

Figure 2: Lineament Proximity Distribution



two predictor variables, whose near linear-separability across classes remain worthy of caution.

Figure 3: Relationship between Antimony and Arsenic



Model Architecture

The logistic regression model, given a dependent variable $Y \in \{0, 1\}$, in our case chosen to be whether a location is within 0.5 km to Gold (denoted `gold_proxim`), and independent variables of the location's Arsenic (`as_level`), Antimony level (`sb_level`), and an indicator of whether this location is within 0.5 km to a lineament (`lineament_proxim`), is defined by equation 1, where a model is constructed based on all three of the predictor variables available in the data set.

In this equation $\eta = log_b(\frac{P(Y=1)}{1-P(Y=1)})$, with $b = e$ in equation 1, though this does not matter. Indeed, should instead one decide on the logistic form $\frac{a^{xb}}{1+a^{xb}}$, since $e^{x \cdot b \cdot ln(a)} = a^{xb}$, one can rewrite $\frac{a^{xb}}{1+a^{xb}} = \frac{e^{x\tilde{b}}}{1+e^{x\tilde{b}}}$, whereby it is shown that this is a matter of scaling. Further, for the purpose of later analyzing the obtained coefficients, exponentiating the log-odds η, allows to recover the odds, as shown in equation 2.

The odds instead of the log-odds allow a more traditional interpretation of the concept of event-probability, by allowing to interpret the ratio of the probability of success over that of failure. These coefficients deviate from their interpretation in the traditional Linear Regression which accompanies additive effects. Instead in the Logistic Regression the exponentiated coefficients indicate the change in odds in multiplicative scale of our regressand given a unit increase of the coefficient-associated predictor. Iteratively reweighted least-squares is employed to find the maximum likelihood estimates of the constructed model, due to both its ability to confront the lack of closed-form solution offered by the logistic regression and its minimal-step convergence resulting from its use of both first and second order derivatives.

Vital assumptions underlying the constructed model which aims to predict a binary outcome, include the independence of errors and observations, absence of multicollinearity, linearity between the logit regressand and continuous variables, lack of strongly influential outliers, and a sufficiently large number of observations per outcome group.

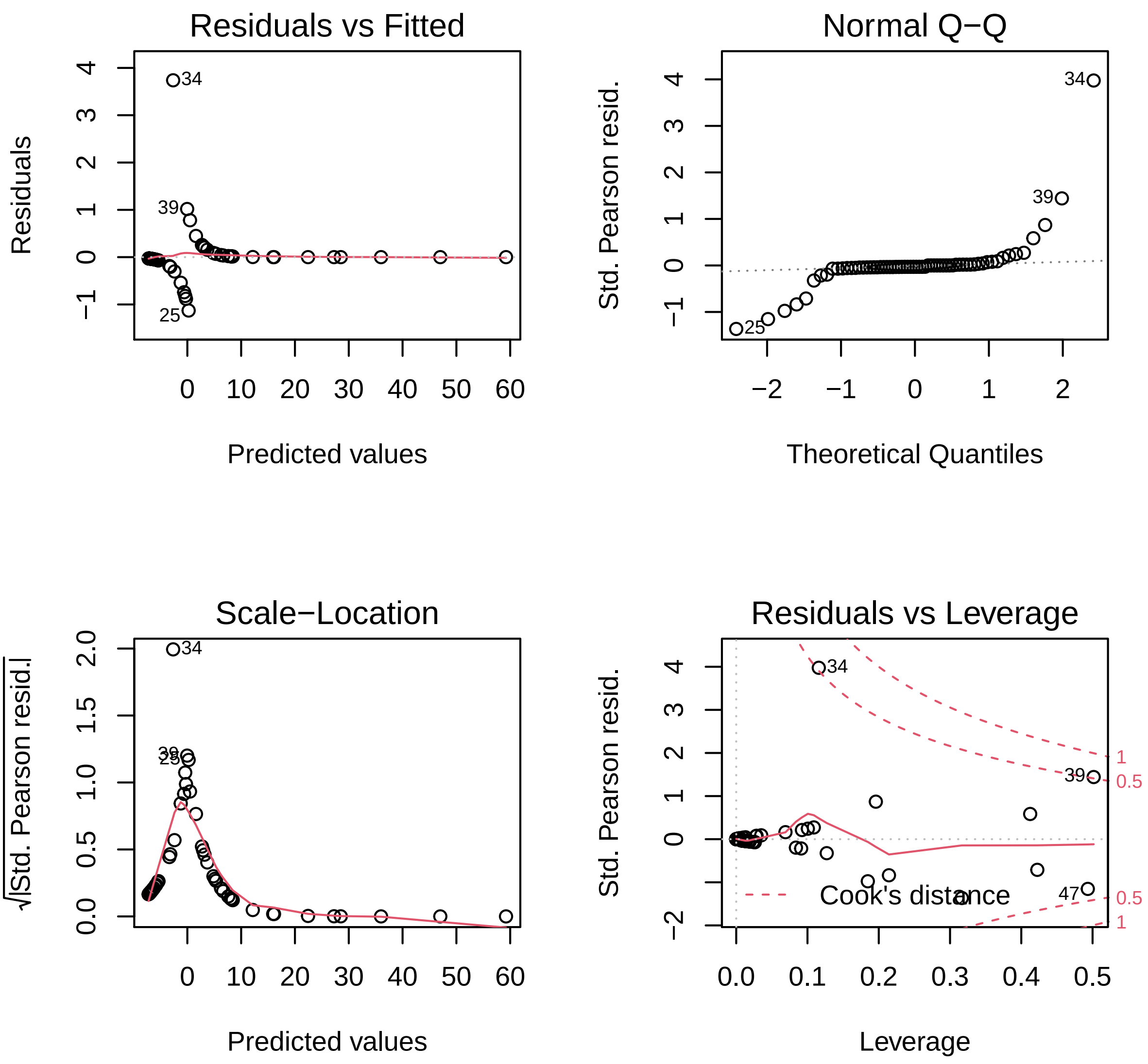
$$\eta = ln(\frac{P(Y=1)}{1-P(Y=1)}) = \widehat{\beta_0} + \widehat{\beta_1} \cdot \texttt{as\_level} + \widehat{\beta_2} \cdot \texttt{sb\_level} + \widehat{\beta_3} \cdot \texttt{lineament\_proxim} \tag{1}$$

$$\frac{P(Y=1)}{1-P(Y=1)} = e^{\widehat{\beta_0} + \widehat{\beta_1} \cdot \texttt{as\_level} + \widehat{\beta_2} \cdot \texttt{sb\_level} + \widehat{\beta_3} \cdot \texttt{lineament\_proxim}} \tag{2}$$

Results & Discussion

The coefficients obtained for the final model, their standard errors, and their exponentiated coefficients are displayed in Table 1. Using the Hosmer and Lemeshow goodness of fit test on the fitted model, one obtains a p-value of 76.23%. This does not necessarily indicate that the model at hand is a good fit, just that there is not sufficient evidence to say that it is poor (recall that the H&L goodness of fit test has as null hypothesis that the model fits the data well). The Akaike information criterion defined by maximizing (minimizing) the (negative) log-likelihood of the given model given a penalty for additional parameters, is given by $AIC = -2/N * LL + 2 * k/N$, with N observations, a log-likelihood of LL and k model parameters. Its low value suggests that model complexity is low, which may otherwise have justified concerns of potential overfitting biasing the Hosmer and Lemeshow goodness of fit test. Further, utilizing Nagelkerke's pseudo-R^2 as a measure of the performance fit and predictive ability, one obtains 0.92, which is remarkably high even for in-sample data, though due to shortcomings and lack of explanatory ability of pseudo-R^2 measures, possible interpretation is somewhat limited. The training procedure was able to classify a total of 61/64 data points correctly, indicating a high level of sample separability between both classes.

Figure 4: Analysis of the residuals of the model presented in Equation 1.



Despite this separability and the model's accuracy, Figure 4 illustrates a hyperbolic shape around the transition point from no-gold locations to those locations with gold. The misclassified points around this cutoff led to modelling attempts with alternative link functions for the generalized linear model, though none attempted (Poisson, Quasibinomial, Gamma) were able to overcome this phenomenon. Figure 5 illustrates the linear relationship of the logit of the predicted outcome variable and the linear predictors, which for the most part suggests that our logistic-linear assumption is upheld, although the upward sloping of the Arsenic levels and larger dispersion around higher levels of Arsenic could hint otherwise (although the small sample size does not aid in making these statements). The presence of a few outliers, determined by Cook's distance, as shown in the fourth quadrant of figure 4, seems problematic in terms of the modelling procedure, as also the normality assumption of the logistic regression's errors is put into question in the second quadrant, with the standardized residuals presenting longer tails than the expected normal distribution.

Figure 5: Relationship of logit predictors and linear outcome

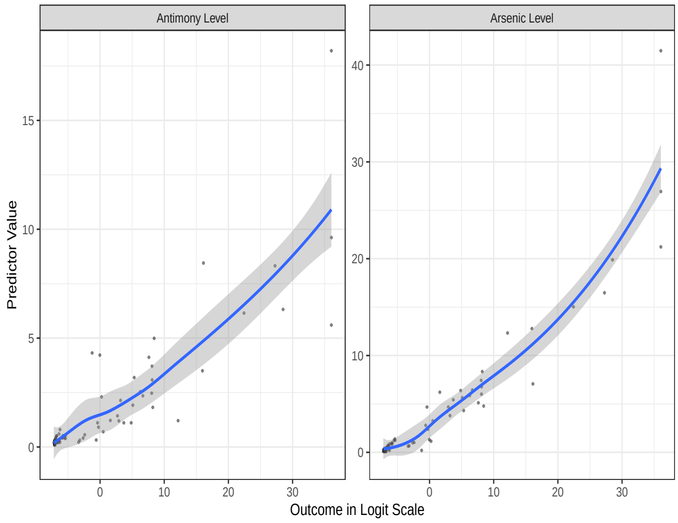


Table 1: Logistic Regression Result

	Dependent variable:
	`gold_proxim`
$\widehat{\beta}_0$	-7.61^{**} (3.17) exp=0.00
$\widehat{\beta}_1$ (`as_lvl`)	1.21^{**} (0.49) exp=3.34
$\widehat{\beta}_2$ (`sb_lvl`)	1.42^{*} (0.73) exp=4.14
$\widehat{\beta}_3$ ($\mathbb{1}_{\texttt{lineament\_proxim1}}$)	3.20^{*} (1.89) exp=24.47
Observations	64
Log Likelihood	−7.10
Akaike Inf. Crit.	22.19
Note:	*p<0.10; **p<0.05; ***p<0.01

In Table 1 is noticeable that out of the intercept and the three included predictor variables, only the intercept and the Arsenic level appear significant on the $\alpha = 5\%$ significance level, while Antimony and Lineament proximity are only significant on the 10% level. The exponentiated intercept indicates the odds of $Y = 1$ given that `as_level`, `sb_level` and `lineament_proxim` are zero. Given the distributions examined in figure 1, these exact values for Arsenic and Antimony levels are unlikely, though the Arsenic and Antimony levels remain moderately close. Then, there would be a mere 5% chance of striking gold in such conditions, which proves a risky discovery project in terms of minimizing discovery costs. As a reminder, the exponentiated coefficients correspond to the multiplicative effect of a unit increase of the regressor upon the regressand, in which case a unit increase of Arsenic and Antimony levels correspond to increasing the odds of gold at a given location by 3.34 and 4.14 respectively. Given their median values of 6.30 and 2.81, this represents a sizable increase in odds of striking gold, enabling companies to optimize their discovery costs should this analysis be replicated on larger-sampled data. However, the proximity of

Lineament is accompanied by on median an even stronger predictive ability in the chance of striking gold, increasing odds by over 24 times.

It remains possible that due to endogenous factors not included in the modelling procedure, Omitted Variable Bias is present, which distorts both coefficient estimates and significance. Indeed it remains largely surprising the insignificance of the factors given $\alpha = 5\%$, given the sample separability they provide. Predictive power of the model appears high while causality remains an undetermined phenomenon for the chosen variables.

Conclusion

In conclusion, this study in a first part presented the geological practice of mineral targeting, which leverages statistical analyses to maximise mineral discovery ability in favor of reducing exploration costs, before presenting a brief examination of the data ($N = 64$) utilized, comprising of Arsenic and Antimony levels, as well as indicators for Lineament to determine Gold proximity. A logistic model is devised and leveraged in order to investigate predictive power of the chosen variables. The model in-sample fit is unsurprisingly, given the near separability of independant variables, highly predictive, though significance in contrast is not given for most at the $\alpha = 5\%$ level. This remains surprising despite the model classifying 61/64 observations correctly and hints at examining, given a larger data set, potentially the exclusion of endogenous alternative factors leading to a case of Omitted Variable Bias.

Supplementary material

The data and code used to replicate the methodology of this study may be found here (https://github.com/KamranSattary/AppBioStats).

References

Bonham-Carter, G. F. (1994). Geographic information systems for geoscientists-modeling with gis. *Computer methods in the geoscientists*, *13*, 398.

Sahoo, N. R., & Pandalai, H. S. (1999). 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.