"Reviewing the review platforms: a privacy perspective" conditions for logistic regression

1 Conditions for logistic regression

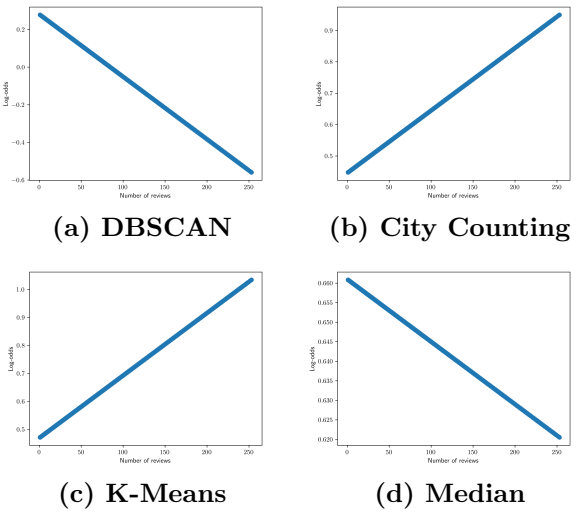
In Section 5.1 of the paper "Reviewing review platforms: a privacy perspective", logistic regression is applied to our data to assess the impact on the identifiability of reviewers based on two key factors, namely the amount of reviews and the spatial entropy of reviews by a reviewer. In order to be able to perform this logistic regression, several conditions need to be met.

- **The dependent variable is binary.** The dependent variable is in our case whether or not the home city of a reviewer is accurately predicted by our methodology. As this is a binary value (yes or no) this requirement is met.
- **The observations are independent.** Reviewers make reviews independently of each other.
- **The relation between the independent variables and log-odds is linear.** The log-odds is the natural logarithm of the odds-ratio and can be defined by the following equation:

$$\text{log-odds} = \ln\left(\frac{p}{1-p}\right)$$

  To prove this condition is met, scatter plots are drawn that present the relation between the independent variable and the log-odds. Figure 1 and 2 demonstrate a clearly linear relation for both the amount of reviews and the spacial entropy respectively. Therefore this condition is also met.

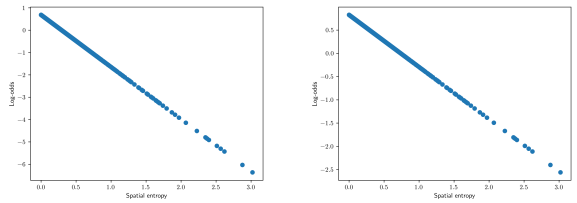
- **No strongly influential outliers.** This condition can be checked by applying the Cook's distance metric [1]. The metric measures the influence of a point when applying regressions. When points are too influential ($Cook'sDistance > 4/N$ with $N =$ the number of observations), they are dropped from the the dataset. In our case, 1.9 to 3.3 % of the data points were dropped.
- **No multicollinearity.** This is not applicable in this specific case, as each logistic regression is only performed with one independent variable.



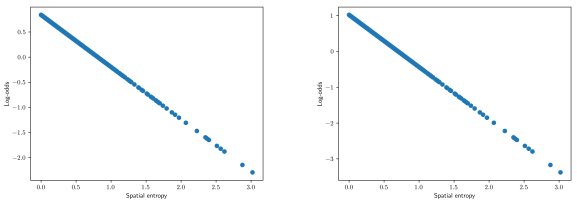
(a) DBSCAN **(b) City Counting**

(c) K-Means **(d) Median**

Figure 1: Log-odds graphs for the amount of reviews

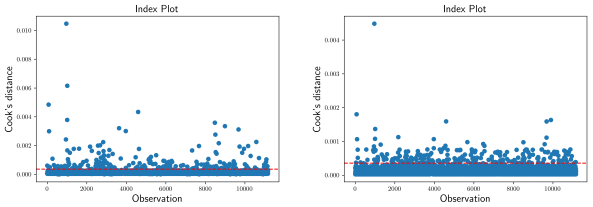


(a) DBSCAN **(b) City Counting**

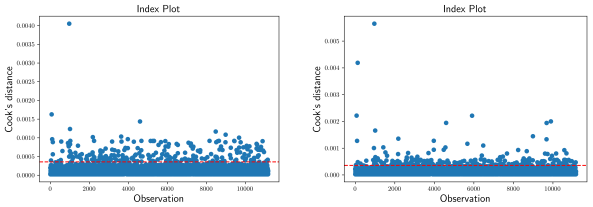


(c) K-Means **(d) Median**

Figure 2: Log-odds graphs for the spatial entropy

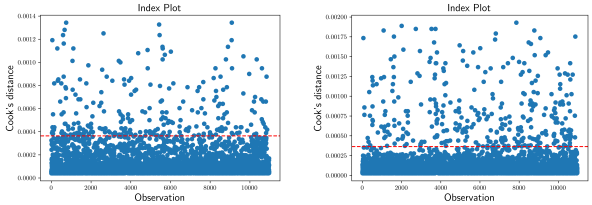


(a) DBSCAN **(b) City Counting**

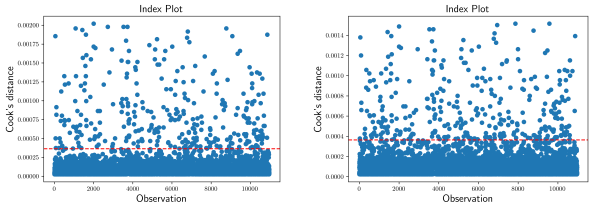


(c) K-Means **(d) Median**

Figure 4: Cook's distance graphs for the spatial entropy



(a) DBSCAN **(b) City Counting**



(c) K-Means **(d) Median**

Figure 3: Cook's distance graphs for the amount of reviews

References

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