Different Slopes for Different Folks

Mining for Exceptional Regression Models with Cook's Distance

Wouter Duivesteijn
LIACS, Leiden University
The Netherlands
wouterd@liacs.nl

Ad Feelders
ICS, Utrecht University
The Netherlands
ad@cs.uu.nl

Arno Knobbe
LIACS, Leiden University
The Netherlands
knobbe@liacs.nl

ABSTRACT

Exceptional Model Mining (EMM) is an exploratory data analysis technique that can be regarded as a generalization of subgroup discovery. In EMM we look for subgroups of the data for which a model fitted to the subgroup differs substantially from the same model fitted to the entire dataset. In this paper we develop methods to mine for exceptional regression models. We propose a measure for the exceptionality of regression models (Cook's distance), and explore the possibilities to avoid having to fit the regression model to each candidate subgroup. The algorithm is evaluated on a number of real life datasets. These datasets are also used to illustrate the results of the algorithm. We find interesting subgroups with deviating models on datasets from several different domains. We also show that under certain circumstances one can forego fitting regression models on up to 40% of the subgroups, and these 40% are the relatively expensive regression models to compute.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

Subgroup Discovery, Exceptional Model Mining, Linear Regression, Cook's Distance

1. INTRODUCTION

Exceptional Model Mining (EMM) [16, 8] is an exploratory data analysis technique that can be regarded as a generalization of subgroup discovery. In subgroup discovery the aim is to find subgroups of the data for which the distribution of a single target variable deviates from its distribution in the entire dataset. In EMM we look for subgroups for which a model fitted to the subgroup differs substantially from the same model fitted to the entire dataset. This is a general framework which can be used for different purposes, depending on the type of model that is fitted, and how one measures the difference between models. In this paper we focus on the "work horse" of data analysis: linear regression models.

Let us consider an example to illustrate why the results of such an analysis might be of interest. The economic *law of demand* states that (all else equal) if the price of a product increases, the demand for the product will decrease. In a regression model this would result in a negative slope when we regress demand on price. However, under specific conditions, people tend to buy more of a product when the price increases [13]. Hence, for those exceptional cases, we would get a positive slope of the regression line. This idea was first published in 1895 [17]. Over one hundred years later, in 2008, this so-called *Giffen behavior* was for the first time observed in a field experiment [13]. In certain theoretically induced subsets (the poor, but not too poor) of households in Hunan, China, demand for rice rose when the price increased. The algorithm we propose is able to find such an exceptional subgroup in the data automatically.

Testing whether the regression coefficients for two groups are different is in fact common practice in applied regression analysis. One common way of doing this is through the use of dummy (i.e. binary) variables. Consider for example the problem of predicting house prices. Suppose we know for each house its selling price, lot size and whether or not it has air conditioning. To test whether the effect of lot size on selling price is different for houses with and without air conditioning, one could estimate the model

$$\text{Price} = \beta_0 + \beta_1 \times \text{Lot Size} + \beta_2 \times \text{Airco} \times \text{Lot Size}, \quad (1)$$

where Airco=1 if the house has air conditioning and Airco=0 otherwise. If we consider these two cases separately we see that the regression equation becomes

$$\text{Price} = \beta_0 + \beta_1 \times \text{Lot Size},$$

if Airco=0, and

$$\text{Price} = \beta_0 + (\beta_1 + \beta_2) \times \text{Lot Size},$$

if Airco=1. Hence, to test whether these two groups of houses have different slopes, we can test whether the coefficient estimate of β_2 in model (1) is significant. In this setup we have to decide in advance for which groups we want to test whether they have different slopes in the regression.

The algorithm presented in this paper can be regarded as a way of automatically finding the subgroups for which the slopes are substantially different. The subgroup description can be more complex than the simple condition in the above

KDD'12, August 12–16, 2012, Beijing, China.

example: it can be a conjunction of conditions on categorical and numeric variables as is common in subgroup discovery. The fact that many textbooks on applied regression analysis (see for example chapter 10 of [14]) devote considerable attention to the problem of *testing* whether two regressions (for different subgroups of the data) are different, underlines the relevance of such questions in practical data analysis and motivates our work. Essentially we embed an old regression technique in a new Exceptional Model Mining setting, to automate a traditional regression problem.

This paper is organized as follows. In Section 2 we introduce some notation, discuss the basic EMM framework, and present the combination of EMM and linear regression models. In Section 3 we propose a quality measure for the exceptionality of subgroups, which measures the distance between the coefficient vector of the global model, and the coefficient vector of the subgroup model. In this section we also consider possibilities to limit the number of models that have to be fitted on subgroups, by using bounds that can be computed quite easily from the data and the global model. In Section 4 we discuss the extent to which we can prune the search space, while in Section 5 we illustrate the application of our algorithm by analyzing a number of publicly available real life datasets. Finally, Section 6 concludes.

2. PRELIMINARIES

Throughout this paper, we assume a dataset $\mathcal{D}$ to be a bag of N *records* $r \in \mathcal{D}$ of the form

$$r = \{a_1, \ldots, a_k, x_1, \ldots, x_{p-1}, y\}$$

where k is a positive integer and p is an integer such that $p \geq 2$. We call $a_1, \ldots, a_k$ the *attributes* of r, and $x_1, \ldots, x_{p-1}, y$ the *targets* of r. Each target is assumed to be numeric, while the attributes are taken from an unrestricted domain $\mathcal{A}$. The requirement that the x_i are numeric may seem very restrictive, but in fact nominal variables can be handled in the usual way by creating "dummy" (binary) variables. We refer to the ith record by r^i.

For our definition of subgroups we need to define *patterns*. These are functions $P : \mathcal{A} \to \{0, 1\}$. A pattern P *covers* a data point r^i if and only if $P\left(a^i\right) = 1$.

Definition (Subgroup). A *subgroup* corresponding to a pattern P is the bag of data points $G_P \subseteq \mathcal{D}$ that P covers:

$$G_P = \left\{ r^i \in \mathcal{D} \;\middle|\; P\left(a^i\right) = 1 \right\}$$

From now on we omit the P if no confusion can arise, and refer to a subgroup as G.

In order to objectively evaluate a candidate pattern in a given dataset, we need to define a *quality measure*. For each pattern P in the pattern language $\mathcal{P}$, this function measures how interesting the model is that we induce on G_P.

Definition (Quality Measure). A *quality measure* is a function $\varphi_{\mathcal{D}} : \mathcal{P} \to \mathbb{R}$ that assigns a unique numeric value to a pattern P, given a dataset $\mathcal{D}$.

2.1 EMM revisited

Exceptional Model Mining [16, 8] is a data mining framework that can be seen as a generalization of the Subgroup Discovery (SD) framework. SD strives to find patterns that satisfy certain user-specified constraints. Usually these constraints include lower bounds on the quality of the pattern ($\varphi(P) \geq lb_1$) and size of the induced subgroup ($|G_P| \geq lb_2$). More constraints may be imposed as the question at hand requires; domain experts may for instance request an upper bound on the complexity of the pattern. Most common SD algorithms traverse[1] the search space of candidate patterns in a general-to-specific way: they treat the space as a lattice whose structure is defined by a *refinement operator* $\rho : \mathcal{P} \to 2^{\mathcal{P}}$. This operator determines how patterns can be extended into more complex patterns by atomic additions. Most applications (including ours) assume ρ to be a *specialization operator*: $\forall\ P_s \in \rho(P_g)\ :\ P_g \succeq P_s$ (i.e. P_s is more specialized than P_g). The algorithm results in a ranked list of patterns (or the corresponding subgroups) that satisfy the user-defined constraints.

In traditional SD there is only a single target variable. Hence, the typical quality measure contains a component indicating how different the distribution over the target variable in the subgroup is, compared to its distribution in the whole dataset. Since unusual distributions are easily achieved in small subsets of the dataset, the typical quality measure also contains a component indicating the size of the subgroup. Thus, whether a pattern is deemed interesting depends on both its exceptionality and the size of the corresponding subgroup.

EMM can be seen as an extension of SD. Rather than the regular single target variable, EMM uses a more complex target concept. For each subgroup under consideration, we induce a model on the targets. Then quality measures are defined that indicate how exceptional the model fitted on the targets in the subgroup is, compared to the model fitted on the targets in the whole dataset. For example, [16] proposes quality measures for correlation models, a simple linear regression model, and classification models.

In the EMM setting, usually the *beam search* strategy is chosen, which performs a level-wise search. On each level, the best w patterns according to our quality measure φ are selected, and refined to create the candidate patterns for the next level. The search is constrained by an upper bound on the complexity of the pattern and a lower bound on the support of the corresponding subgroup. This search strategy combines the advantages of a greedy method with those of the implicit parallel search: as on each level w alternatives are considered, the search process is less likely to end up in a local optimum than a pure greedy approach, but the selection of the w best patterns at each level keeps the process focused and thus more tractable.

There are Subgroup Discovery techniques that exhaustively explore the search space. These however usually either compel the attributes to be nominal [9, 15] or impose an anti-monotonicity constraint on the quality measure [11]. Since we do not want such restrictions on the attributes or the quality measure, we employ heuristic search.

[1] we consider the exact search strategy to be a parameter of the algorithm

2.2 Linear Regression Model

The previous section introduced Exceptional Model Mining in its general form. In this paper, however, we are concerned with one particular instance: the linear regression model:

$$Y = X\beta + \varepsilon$$

where Y is the $N \times 1$ vector of y-values from our dataset, X is the $N \times p$ full rank matrix of which the first column consists of N times the value 1 and column $i+1$ contains the x_i-values from our dataset, β is the unknown $p \times 1$ vector consisting of the regression parameters, and ε is the $N \times 1$ vector of randomly distributed errors such that $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I$. Of course, I denotes the $N \times N$ identity matrix.

Given an *estimate* of the vector β, denoted $\hat{\beta}$, one can compute the vector of *fitted values* $\hat{Y}$. These quantities can be used to assess the appropriateness of the fitted model, by looking at the *residuals* $e = Y - \hat{Y}$. We will estimate β with the ordinary least squares method, which minimizes the sum of squared residuals. This leads to the estimate:

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y$$

After computing the vector of fitted values, we find that we can now write the corresponding residual vector as:

$$e = (e_i) = Y - \hat{Y} = \left(I - X\left(X^\top X\right)^{-1} X^\top\right) Y$$

We will denote a part of this equation by V:

$$V = (v_{ij}) = X\left(X^\top X\right)^{-1} X^\top$$

This matrix was dubbed the *hat matrix* by John W. Tukey, since $\hat{Y} = VY$, i.e. the hat matrix transforms Y into $\hat{Y}$ [12].

2.3 Regression meets EMM

To mine for exceptional regression models, we have to come up with a good quality measure. It stands to reason that this quality measure should quantify the difference between the coefficient vector $\hat{\beta}$ estimated on the dataset, and the vector $\hat{\beta}^G$ estimated on the subgroup. One could for example use the squared Euclidian distance

$$(\hat{\beta}^G - \hat{\beta})^\top (\hat{\beta}^G - \hat{\beta}) = \sum_{i=1}^{p} (\hat{\beta}_i^G - \hat{\beta}_i)^2 \qquad (2)$$

to measure the quality of subgroup G. One can argue that in many applications we are not really interested in the influence of all variables on y, but just in the influence of one, or a small subset, of them. The other variables are merely included in the regression to obtain good estimates of the coefficients we are interested in. This can be easily accommodated by summing over the subset of interest in Equation (2), were one so inclined.

As Equation (2) suggests, to compute the quality of a subgroup we have to fit a model on it in order to obtain the estimates $\hat{\beta}^G$. Since one has to evaluate many subgroups, this can be computationally quite demanding. Therefore, it is of some interest to determine whether such explicit computation can be avoided or limited.

In the next section we propose a more sophisticated quality measure for exceptional regression models, and look at the possibilities of limiting explicit model fitting on subgroups.

3. COOK'S DISTANCE

In the previous section, we suggested the squared Euclidian distance between estimated coefficient vectors as a quality measure. The disadvantage of this measure is that it ignores the variance of the estimator $\hat{\beta}$, and the covariances between $\hat{\beta}_i$ and $\hat{\beta}_j$. For example, if $\hat{\beta}_i$ has a large variance compared to $\hat{\beta}_j$, then a given change in $\hat{\beta}_i$ should contribute less to the overall quality than the same change in $\hat{\beta}_j$, because the change in $\hat{\beta}_i$ is more likely to be caused by random variation. This suggest that

$$(\hat{\beta}^G - \hat{\beta})^\top [\mathrm{Cov}(\hat{\beta})]^{-1} (\hat{\beta}^G - \hat{\beta})$$

might be a better distance measure than the normal Euclidian distance. In fact this expression is equivalent to Cook's distance up to a constant scale factor. R. Dennis Cook originally introduced his distance [3] in 1977 for determining the contribution of single records to $\hat{\beta}$. In this section we discuss this distance measure in detail.

3.1 Cook's distance for single observations

Recall that the least squares estimate of β is

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y.$$

The corresponding residual vector becomes

$$e = Y - \hat{Y} = Y - X\hat{\beta} = (I - V)\, Y,$$

where V is the hat matrix defined in Section 2.2. The covariance matrices of $\hat{Y}$ and e, respectively, are

$$\mathrm{Var}\left(\hat{Y}\right) = V\sigma^2, \qquad \mathrm{Var}\,(e) = (I - V)\,\sigma^2$$

Finally, Cook states that according to normal theory [10], the $(1-\alpha) \times 100\%$ confidence ellipsoid for the unknown vector, β, is given by the set of all vectors β^* satisfying

$$\frac{\left(\beta^* - \hat{\beta}\right)^\top \widehat{[\mathrm{Cov}(\hat{\beta})]^{-1}} \left(\beta^* - \hat{\beta}\right)}{p} =$$

$$\frac{\left(\beta^* - \hat{\beta}\right)^\top X^\top X \left(\beta^* - \hat{\beta}\right)}{ps^2} \leq F(p, N-p, 1-\alpha)$$

where

$$s^2 = \frac{e^\top e}{N-p}, \qquad \widehat{\mathrm{Cov}}(\hat{\beta}) = s^2\left(X^\top X\right)^{-1}$$

and $F(p, N-p, 1-\alpha)$ is the $1-\alpha$ probability point of the central F-distribution with p and $N-p$ degrees of freedom. Here, s^2 is the unbiased estimator for σ^2.

Now the stage has been set to determine the degree of influence of single records. Suppose we want to know how record r^i influences $\hat{\beta}$. Then one could naturally compute the least squares estimate for β with the record removed from the dataset. Let us denote this estimate by $\hat{\beta}_{(i)}$. We can adapt the confidence ellipsoid as an easily interpretable measure of the distance between $\hat{\beta}_{(i)}$ and $\hat{\beta}$. Hence, *Cook's distance* is defined as:

$$D_i = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)^\top X^\top X \left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{ps^2} \qquad (3)$$

Suppose for example that for a certain record r^i we find that $D_i \approx F(p, N-p, 0.5)$. Then removing r^i moves the

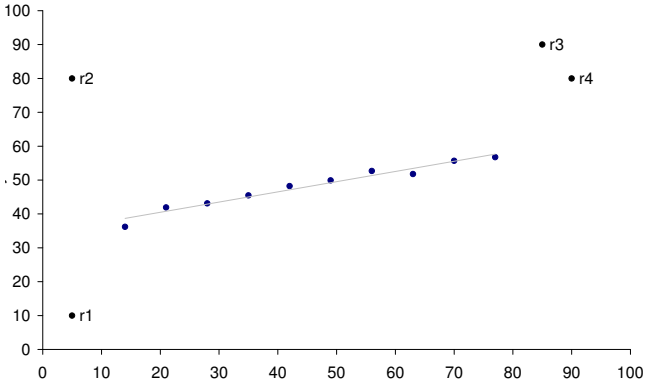

Figure 1: Records r^1 and r^2 are individually influential, but not jointly. Conversely, r^3 and r^4 are jointly influential, but not individually.

least squares estimate to the edge of the 50% confidence region for β based on $\hat{\beta}$.

Bingham [2] showed that Equation (3) can be rewritten in the form

$$D_i = \frac{\left(\hat{Y}_{(i)} - \hat{Y}\right)^\top \left(\hat{Y}_{(i)} - \hat{Y}\right)}{ps^2}$$

which suggests that apart from the scale factor ps^2, D_i is the ordinary squared Euclidean distance that the fitted vector moves when r^i is removed from the dataset.

Cook's distance would not be particularly useful if one really would have to compute $\hat{\beta}_{(i)}$ for each record in the dataset. However, Cook shows [3] that under some mild assumptions D_i can be rewritten as

$$D_i = \frac{t_i^2}{p} \cdot \frac{\mathrm{Var}\left(\hat{Y}_i\right)}{\mathrm{Var}\left(e_i\right)} \tag{4}$$

where t_i is the studentized residual of the ith record. These quantities all relate to the full dataset. Clearly t_i is a measure of how well r^i can be considered an outlier from the assumed model. The ratio of variances measures the relative sensitivity of $\hat{\beta}$ to potential outliers at each data point. Combining these factors hence produces a measure of the impact of any single point on the least squares estimate.

3.2 Cook's distance for multiple observations

Whenever we want to compute the influence of deleting several points simultaneously, as is the case in EMM, one cannot simply use Equation (3) and sum over all records concerned. We will illustrate why with a simple constructed example [4]. Consider linear regression on the dataset from Figure 1. Suppose that we consider removing records r^1 and r^2 from the dataset. If we would remove either of these records, this will have a rather large influence on the slope of the resulting regression line, hence according to Equation (3) both D_1 and D_2 will be large. However, when we remove both records from the dataset, the influences of the records will cancel each other out, and the slope of the regression line will barely change at all: r^1 and r^2 are not *jointly influential*. On the contrary, when removing either record r^3 or r^4 from the dataset, the slope of the regression line will barely change, hence according to Equation (3) both D_3 and D_4 are small. However, their joint influence is quite large: removing both records from the dataset will significantly influence the slope of the regression line.

Hence, we cannot give a reliable measure for the joint influence of a set of records by aggregating over values of D_i. Therefore, Cook and Weisberg extended Cook's distance to cope with deleting multiple records simultaneously [4]. Let I be a vector of indices that specify the m records to be deleted. From now on, we let the subscript (I) denote "with the m cases indexed by I deleted", while the subscript I without parentheses denotes "with only the m cases indexed by I remaining". The only notation that deviates from this rule of thumb is the definition of Cook's distance, which is easily extended to multiple observations:

$$D_I = \frac{\left(\hat{\beta}_{(I)} - \hat{\beta}\right)^\top X^\top X \left(\hat{\beta}_{(I)} - \hat{\beta}\right)}{ps^2} \tag{5}$$

and its geometric interpretation is identical to the geometrical interpretation of D_i. Any subset that has a large joint influence on the estimation of β corresponds to a large D_I.

The fact that the definition of Cook's distance does not follow the notational rule of thumb can be very confusing. We choose to retain the definition in this form to make our work compatible with previously released papers and books. However, it is important to stress the notational anomaly: whenever we write D_I, Cook's distance is computed for the case where the records indexed by I are *deleted*. Whenever we write *anything else* with a subscript I, it is computed for the case where the records indexed by I are *retained*, and all other records are deleted.

Unfortunately, unlike in the case where we deleted only one record, D_I cannot simply be rewritten in a form resembling Equation (4). Hence we need another solution to the problem that computing $\hat{\beta}_{(I)}$ for each candidate subgroup is computationally very expensive.

The upper bounds for Cook's distance are derived [5, pp. 136] by rewriting the numerator of the right hand side of Equation (5) in terms of e_I and V_I. Then the spectral decomposition of V_I is used, rewriting the sub-matrix of the hat matrix in terms of its eigenvalues and eigenvectors. We denote those eigenvalues by $\lambda_1, \ldots, \lambda_m$, and can assume without loss of generality that $0 \leq \lambda_1 \leq \ldots \leq \lambda_m \leq 1$. Notice that if the last inequality is not strict, i.e. $\lambda_m = 1$, then removing the records indexed by I would lead to a rank deficient model, and we cannot properly perform the linear regression. Finally, a proper approximation for these λ_i is required; Cook proposes to use $\mathrm{tr}\left(V_I\right)$ here, but notes that this is only a good approximation under the condition that $\mathrm{tr}\left(V_I\right) < 1$. Assuming that this condition holds, we can bound D_I by:

$$D_I \leq \frac{\mathrm{tr}\left(V_I\right)}{\left(1 - \mathrm{tr}\left(V_I\right)\right)^2} \cdot \frac{\sum_{i \in I} e_i^2}{ps^2} \tag{6}$$

Unfortunately, this bound is potentially different for each I. Cook also gives bounds that hold for all subsets I of a fixed size m. When we fix m and let I vary over all such subsets, we can either use $R^2 = \max_I \left(\sum_{i \in I} e_i^2\right)$, which turns Equation (6) into:

$$D_I \leq \frac{\mathrm{tr}\left(V_I\right)}{\left(1 - \mathrm{tr}\left(V_I\right)\right)^2} \cdot \frac{R^2}{ps^2} \tag{7}$$

or we could use $T = \max_I \left(\sum_{i \in I} v_{ii}\right)$, which turns Equation (6) into:

$$D_I \leq \frac{T}{(1-T)^2} \cdot \frac{\sum_{i \in I} e_i^2}{ps^2} \tag{8}$$

Both simplifications R^2 and T can be combined to turn Equation (6) into:

$$D_I \leq \frac{T}{(1-T)^2} \cdot \frac{R^2}{ps^2} \tag{9}$$

Rather obviously, there are relations between the bounds, i.e. (6) ≤ (7) ≤ (9) and (6) ≤ (8) ≤ (9).

3.3 Subsets of $\hat{\beta}$

For practical purposes one might not be interested in computing Cook's distance based on the entire parameter vector $\hat{\beta}$. For instance, one might be interested in the influence records have on the regression coefficient corresponding to one particular attribute, while excluding the intercept and other coefficients from the evaluation. To this end, Cook and Weisberg [5] introduce the zero/one-matrix Z, with dimensions $q \times p$, where q is the number of elements of $\hat{\beta}$ that we are interested in (hence $q \leq p$). The matrix Z is defined in such a way that $\psi = Z\beta$ are the coefficients of interest. Hence, if we are interested in the last q elements of $\hat{\beta}$, Z will start from the left with $p - q$ columns containing all zeroes, followed by a $q \times q$ identity matrix ($Z = (\mathbf{0}, \mathbf{I_q})$).

When using this transformation, Cook's distance (Equation (5)) becomes:

$$D_I^{\psi} = \frac{\left(\hat{\beta}_{(I)} - \hat{\beta}\right)^{\top} Z^{\top} \left(Z \left(X^{\top} X\right)^{-1} Z^{\top}\right)^{-1} Z \left(\hat{\beta}_{(I)} - \hat{\beta}\right)}{qs^2}$$

One can show that $qD_I^{\psi} \leq pD_I$ for all I, hence one can make all bounds (6)–(9) relevant for D_I^{ψ} by multiplying them by the factor $\frac{p}{q}$.

3.4 Cook's distance in EMM

Since Cook's distance is invariant to changes in scale of the variables involved [3], it would make an excellent quality measure for use in EMM:

Definition ($\varphi_{\mathbf{Cook}}$). Let G_P be a subgroup. Its *quality according to Cook's distance* is given by:

$$\varphi_{\text{Cook}}(G_P) = D_I^{\psi}, \text{ where } I = \left\{ i \middle| r^i \in \mathcal{D}, P\left(a^i\right) = 0 \right\}$$

Hence, Cook's distance of a subgroup is the distance bridged when the records that are not covered by the subgroup are simultaneously discarded. This definition seems a bit convoluted; it is constructed in such a way that the notational anomaly discussed in Section 3.2 is repaired: whenever we write $\varphi_{\text{Cook}}(G_p)$, Cook's distance is computed for the case where the records belonging to the subgroup G_p are *retained*.

3.5 Pruning with Cook's bounds

Whenever one has the possibility to enumerate all candidate subgroups for mining with Cook's distance, the bounds (6)–(9) can be used for pruning. In combination with the beam search strategy, we propose to do this in the following way.

Per search level, we determine the number of subgroups S we are interested in retaining. We enumerate all candidate subgroups in descending ordered according to one of the bounds. Then we consider the subgroups in this order.

For each subgroup, we compute the bounds in order of decreasing ease of computation, i.e. first bound (9), then bound (8), then bound (7), and finally bound (6). We check whether any of these bounds has a value that is lower than Cook's distance for the S^{th} best evaluated subgroup so far. If so, we know that Cook's distance for this new subgroup can not enter the top-S, since the bound is an upper bound for Cook's distance. Hence we can skip computing Cook's distance for this subgroup, which saves us the computation of a relatively expensive regression. If none of the bounds help us out, we compute Cook's distance for the new subgroup.

Dataset	N	k	p
Ames Housing	2930	77	3
Auction	1225	3	7
EAEF	2714	32	3
Giffen Behavior	1254	6	16
PC486	6259	3	7
Wine	5000	6	4

Table 1: Some elementary properties of the datasets. N is the total number of records, k is the number of attributes that can be used to define subgroups, and p is the number of coefficients in the fitted regression model.

4. BOUND BEHAVIOR AND PRUNING

Table 1 lists the datasets we have used in the experiments. All datasets are publicly available. The Giffen behavior dataset[2] was used for a study that claimed to provide the first real-world evidence of Giffen behavior, i.e. an upward sloping demand curve [13]. The EAEF dataset[3] was analyzed in [7]. The Wine data was analyzed in [6], and is available from the data archive of the Journal of Applied Econometrics[4]. The Ames Housing data is available from the Journal of Statistics Education data archive[5]. Finally, the PC486 data and the Auction data were analyzed in [19] and [18] respectively, and are both available from the data archive of the Journal of Applied Econometrics.

To illustrate what can reasonably be expected from pruning with the bounds, we simulated their behavior on random subgroups of the EAEF dataset. For each possible subgroup size, we drew a random sample of the data with that size. Then we computed the values of the bounds for these subgroups, when fitting the model

$$\text{Earnings } = \beta_0 + \beta_1 \times \text{YrsOfSchool}$$

The results can be found in Figure 2. The figure depicts the subgroup size on the x-axis (linear scale), and the values of the bounds on the y-axis (logarithmic scale).

[2]The data can be downloaded from: `www.aeaweb.org/articles.php?doi=10.1257/aer.98.4.1553`

[3]The data can be obtained from `www.oxfordtextbooks.co.uk/orc/dougherty4e/`

[4]`http://econ.queensu.ca/jae/`

[5]`http://www.amstat.org/publications/jse/jse_data_archive.htm`

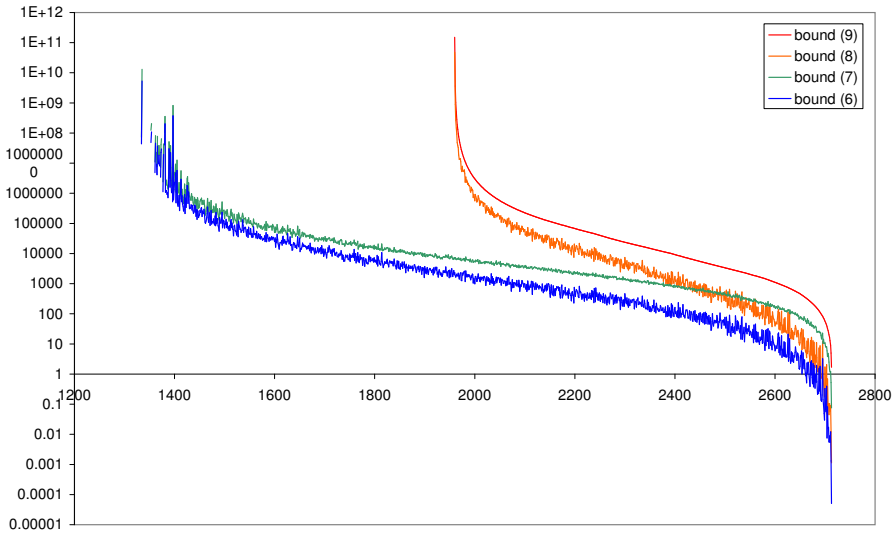

Figure 2: Bound values (logarithmic scale) for random subgroups of different sizes on the EAEF dataset. Fitted model: Earnings $= \beta_0 + \beta_1 \times$ **YrsOfSchool.**

The EAEF dataset has 2714 records, so when a subgroup approaches this size it will correspond to deleting very few records, and as one would expect Cook's distance becomes very small, as do the bounds. Furthermore, one notices that the bound quality lines do not extend all the way to subgroup size 0. This is caused by limitations in the approximations used in the bounds. As we mentioned in Section 3.2, the bounds are only good approximations whenever $\mathrm{tr}\,(V_I) < 1$. When this constraint is not satisfied, the bounds cannot be computed. For bounds (8) and (9), the quantity T is used as an estimate for $\mathrm{tr}\,(V_I)$, but this too only makes sense when $T < 1$, or else the bounds cannot be computed.

The practical upshot is that for subgroups smaller than 1960 records, bounds (8) and (9) cannot be computed. For subgroups smaller than roughly 1250 records, this also holds for bounds (6) and (7). When viewed as a percentage of the number of records in the datasets, we find that these borders are roughly the same over all datasets: bounds (6) and (7) can only be computed when the subgroup contains at least 50% of the records, and bounds (8) and (9) only when the subgroup contains at least 75% of the records. We also find that the more complex the model we fit, the further these thresholds move towards larger percentages.

The bounds can not be computed for at least half of the subgroups we consider, and the bound values tend to increase enormously just before these threshold values are reached. However, the bounds are computable for the largest subgroups, and the computation of the hat matrix is quadratic in the subgroup size. Hence whenever we can prune a subgroup, it always takes a relatively expensive regression computation out of the total runtime.

4.1 Empirical bound evaluation

To empirically see how the bounds perform, we performed a depth-1 subgroup discovery run on each dataset, with the goal to find the top-1 subgroup. When numeric attributes were used to generate candidate subgroups, we split them into 12 equal-sized bins. We discarded any subgroup that covered less than 100 records, since we consider these too small to be considered interesting from a statistical point of view. For each bound we counted how often it was computed, and how often it caused a subgroup to be pruned.

The results can be found in Table 2. This table features the datasets, dataset characteristics, number of times every bound is computed, number of subgroups pruned with every bound, fraction of candidate subgroups for which at least one bound was computable, and fraction of candidate subgroups that were pruned. Notice that there is a strong dependency between the "Bound computed" and "Subgroups pruned" columns: in the Ames Housing dataset we can compute bound (9) for 196 subgroups, of which we can prune 155, so only 41 subgroups remain for which we compute bound (8). However, the number of subgroups for which we compute bound (7) is larger, since the condition under which this bound is computable is less strict than the condition for bound (8) and (9). Of the 228 subgroups for which we compute bound (7) we can prune 191, leaving 37 subgroups for which we compute bound (6).

As we indicated in the previous section, the fraction of subgroups for which we can compute the bounds is strongly dependent on the complexity of the fitted model. As we can see from the table, in the datasets for which $3 \leq p \leq 4$ we can compute bounds for over 40% of the subgroups, in the datasets for which $p = 7$ we can compute bounds for $33 - 35\%$ of the subgroups, and in the dataset for which $p = 16$ we can compute bounds for just 1% of the subgroups. This dependency becomes somewhat less direct when we look at the percentage of subgroups we can actually prune, since this is relatively low for the EAEF dataset on which we fit a relatively simple model. However, apart from this one dataset, we still see a strong relation between model simplicity and pruning success.

Since we are rarely interested in only the one best-performing subgroup, we replicate these experiments with the goal to find the top-50 subgroups. Since we need to have considered at least 50 subgroups before we can make sure others

Dataset	N	$\lvert\mathcal{C}\rvert$	p	Bounds computed				Subgroups pruned				$\frac{\lvert\text{bounded } \mathcal{C}\rvert}{\lvert C\rvert}$	$\frac{\lvert\text{pruned } \mathcal{C}\rvert}{\lvert\mathcal{C}\rvert}$
				(9)	(8)	(7)	(6)	(9)	(8)	(7)	(6)		
Ames Housing	2930	980	3	196	41	228	37	155	28	191	11	0.419	0.393
Auction	1225	40	7	5	0	9	5	5	0	4	0	0.350	0.225
EAEF	2714	204	3	35	29	68	68	6	9	0	21	0.407	0.176
Giffen Behavior	1254	100	16	1	1	1	1	0	0	0	1	0.010	0.010
PC486	6259	6	7	0	0	2	1	0	0	1	0	0.333	0.167
Wine	5000	56	4	2	2	26	20	0	0	6	11	0.464	0.304

Table 2: Pruning results for depth-1 EMM runs, when looking for the top-1 subgroup. N is the total number of records, $\mathcal{C}$ is the set of candidate subgroups considered, and p is the number of coefficients in the fitted regression model.

Dataset	N	$\lvert\mathcal{C}\rvert$	p	Bounds computed				Subgroups pruned				$\frac{\lvert\text{bounded } \mathcal{C}\rvert}{\lvert C\rvert}$	$\frac{\lvert\text{pruned } \mathcal{C}\rvert}{\lvert\mathcal{C}\rvert}$
				(9)	(8)	(7)	(6)	(9)	(8)	(7)	(6)		
Ames Housing	2930	980	3	196	125	272	122	71	68	150	44	0.419	0.340
EAEF	2714	204	3	35	34	77	77	1	5	0	11	0.407	0.083

Table 3: Pruning results for depth-1 EMM runs, when looking for the top-50 subgroups.

will not enter the top-50 based on their bounds, we know in advance that there will be little or no pruning possible for the Auction, PC486, and Wine datasets. We also expect to gain little information from the Giffen Behavior dataset, hence Table 3 encompasses the results of these experiments on merely the Ames Housing and EAEF dataset. Notice that the fraction of subgroups we can prune on the Ames Housing dataset has only decreased slightly, while the fraction of subgroups we can prune on the EAEF dataset is cut in half.

We repeated all these experiments with depth-2 subgroup discovery runs with beam width 10. We find that in these experiments, we can barely compute bounds for any level-2 subgroups, let alone prune subgroups. This is caused by the fact that level-2 subgroups are refinements of well-scoring level-1 subgroups, which are usually relatively small. Well-scoring level-1 subgroups almost never cover more than 50% of the records, hence their refinements also almost never do so. Fortunately, that also means that the regression computations for these subgroups is relatively cheap.

5. ILLUSTRATIVE RESULTS

In this section we give a number of examples of the types of results that can be obtained with our algorithm. The reader should keep in mind that this type of analysis should normally be performed in collaboration with a subject area expert who could aid in the interpretation of the results.

5.1 Giffen Behavior Data

This dataset was used for a study that claimed to provide the first real-world evidence of Giffen behavior, i.e. an upward sloping demand curve [13]. As common sense suggests, the demand for a product will usually decrease as its price increases. According to economic textbooks, there are circumstances however, for which we should expect to see an upward sloping demand curve. The common example is that of poor families that spend most of their income on a relatively inexpensive staple food (e.g. rice or wheat) and a small part on a more expensive type of food (e.g. meat). If the price of the staple food rises, people can no longer afford to supplement their diet with the more expensive food, and must consume more of the staple food.

The dataset we analyze was collected in different counties in the Chinese province Hunan, where rice is the staple food. The price changes were brought about by giving vouchers to randomly selected households to subsidize their purchase of rice. The global model estimated in [13] is:

$$\%\Delta\text{staple}_{i,t} = \alpha + \beta\%\Delta p_{i,t} + \sum \gamma\%\Delta Z_{i,t} + \\ + \sum \delta\text{County} \times \text{Time}_{i,t} + \Delta\varepsilon_{i,t},$$

where $\%\Delta\text{staple}_{i,t}$ denotes the percent change in household i's consumption of rice, $\%\Delta p_{i,t}$ is the percent change in the price of rice due to the subsidy (negative for $t = 2$ and positive for $t = 3$), and $\%\Delta Z_{i,t}$ is a vector of percent changes in other control variables including income and household size. County $\times$ Time denotes a set of dummy variables included to control for any county-level factors that change over time. For each household, two changes are observed: the change between periods 2 and 1 ($t = 2$), capturing the effect of giving the subsidy; and the change between periods 3 and 2 ($t = 3$) capturing the effect of removing the subsidy. For further details about the design of the study and the estimation strategy, we refer to [13].

The coefficient of primary interest is β. If $\beta > 0$ we observe Giffen behavior. The other variables are included in the model to control for other possible influences on demand, so that the effect of price can be reliably estimated. Therefore it makes sense to restrict our quality measure to the coefficient β, that is, the quality of a subgroup is proportional to the absolute difference between $\hat{\beta}$ and $\hat{\beta}^G$.

The authors of [13] suggest that for the extremely poor, one might not observe Giffen behavior, because they consumed rice almost exclusively anyway, and therefore have no other possibility than buying less of it in case of a price increase. The Initial Staple Calorie Share (ISCS) was also measured in the study, and the hypothesis is that families with a high value for this variable do not display Giffen behavior. The authors of [13] tried different manually selected thresholds on ISCS; for example, for the subgroup

of households with ISCS > 0.8, indeed it is observed that $\hat{\beta} = -0.585$ (no Giffen behavior) whereas for ISCS ≤ 0.8 they get $\hat{\beta} = 0.466$ (Giffen behavior).

We analyzed this dataset with ISCS as one of the variables on which the subgroups could be defined. At depth 1, the best subgroup we found was ISCS ≥ 0.87 with $\hat{\beta} = -0.96$ (against $\hat{\beta} = 0.22$ for the complete dataset). The size of this subgroup is $n = 106$. This confirms the conclusion that Giffen behavior does not occur for families that almost exclusively consume rice anyway. This conclusion can also be reached by defining subgroups on *income per capita* rather than ISCS. Particularly illustrative examples are the 4^{th}-ranked subgroup: Income per Capita ≤ 64.67, with a slope of -0.41, and the 6^{th}-ranked subgroup: Income per Capita ≥ 803.75, with a slope of 0.79 (strong Giffen behavior).

5.2 EAEF Data

This dataset has been extracted from the National Longitudinal Survey of Youth 1979-(NLSY79). It contains information about hourly earnings of men and women, and information about, among others, their education. For more details, we refer to Appendix B of [7]. We fit a model relating years of schooling and years of work experience to earnings. The model fitted on the complete dataset is:

$$\text{Earnings} = -29.15 + 2.78 \times \text{YrsOfSchool} + 0.63 \times \text{YrsWorkExp}$$

All coefficients in this model are significant at $\alpha = 0.01$, and R^2 is approximately equal to 20%.

The 4th ranked subgroup we found was COLLBARG $= 1$, meaning that the pay was set by collective bargaining. The fitted model for this subgroup of size $n = 533$ is:

$$\text{Earnings} = -8.93 + 1.57 \times \text{YrsOfSchool} + 0.43 \times \text{YrsWorkExp}$$

This suggests that for this group an extra year of schooling on average leads to an increase of just \$1.57 in hourly earnings, compared to \$2.78 for the whole dataset. The same is true for the influence of an extra year of work experience: just \$0.43 for the collective bargaining subgroup, against \$0.63 in the complete dataset. This is consistent with the finding that unions tend to equalize the income distribution, especially between skilled and unskilled workers [1].

5.3 Wine Data

This dataset was analyzed in [6]. It is composed of 9600 observations derived from 10 years (1991-2000) of tasting ratings reported in the Wine Spectator Magazine (online version) for California and Washington red wines. Our analysis uses a random sample of size 5000 from the original data. For a detailed description of the data we refer to [6]. The global model is

$$\text{Price} = -186.61 - 0.0002 \times \text{Cases} + 2.35 \times \text{Score} + 5.51 \times \text{Age},$$

where *Price* is the retail price suggested by the winery, *Score* is the score from the Wine Spectator, *Age* is the years of aging before commercialization, and *Cases* is the number of cases produced (in thousands). All coefficients are significant at $\alpha = 0.01$, and R^2 is approximately equal to 31%. Furthermore, all coefficients have the sign that one would expect based on common sense.

The most deviating subgroup is *Variety = Non-varietal* (alternatives are Pinot noir, Cabernet, Merlot, Zinfandel and Syrah). The regression model for this subgroup is:

$$\text{Price} = -341.92 - 0.0004 \times \text{Cases} + 4.16 \times \text{Score} + 7.22 \times \text{Age}$$

Non-varietal actually means that multiple varieties of grapes are used, and on average these wines are more expensive than the single-variety wines (average price of \$44,16 against \$28,89). People buying those more expensive wines tend to be better informed (e.g. read Wine Spectator Magazine) than the average buyer. This explains to a certain extent why the price of those more expensive wines is more sensitive to its score and age: they have more critical buyers.

5.4 Ames Housing Data

This dataset contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, Iowa from 2006 to 2010. It consists of 2930 observations on 82 variables. The global model is

$$\text{Price} = -108225.05 + 1.93 \times \text{Lot Area} + 44201.87 \times \text{Quality},$$

where *Price* is the sales price of the house in dollars, *Lot Area* is the lot size in square feet, and *Quality* rates the overall material and finish of the house on a scale from 1 to 10. All coefficients are significant at $\alpha = 0.01$, and R^2 is about 67%.

By far the most deviating subgroup we find is where the building type is a *townhouse inside unit*:

$$\text{Price} = -17674.20 + 24.62 \times \text{Lot Area} + 15786.88 \times \text{Quality}$$

The size of this subgroup is $n = 101$, and the R^2 of the model fitted to this subgroup is about 71%. The dependence of price on lot area is much stronger for town houses, whereas the dependence of price on overall quality is less strong than in general. In an attempt to explain this pattern, we note that the average lot area of town houses (2353 square feet) is much smaller than the overall average (10148 square feet) which is largely determined by the predominant building type *single family detached*. Furthermore, it stands to reason that for townhouses a larger part of the lot area is actually occupied by the house itself than for the single family detached houses. This is consistent with a much stronger dependence of their price on the lot area.

6. CONCLUSIONS

In this paper, we have proposed to use Cook's distance in an Exceptional Model Mining setting. This allows us to find subgroups of the data, for which a regression model fitted on certain dedicated target variables is substantially different from that model for the whole dataset. The use of Cook's distance has two large benefits.

On the one hand, Cook's distance has some desirable properties. It is invariant under changes in the scale of a variable, and it explicitly takes the covariance matrix of $\hat{\beta}$ into account. Hence when using Cook's distance we need not worry whether the outcome of the EMM algorithm is influenced by the scale ones attributes happen to arrive in (attributes need not be normalised), or the interactions that happen to be present between the regression parameters.

On the other hand, there are some theoretical upper bounds on Cook's distance, that can be computed without actually performing the relatively expensive regression computations. As we have seen, these bounds can only be computed under certain constraints, which correspond to the subgroup covering at least 50% of the records. On the one hand, this means that we can compute the bounds for relatively few subgroups, but on the other hand, whenever we can

prune a subgroup, we always prune a relatively expensive regression computation, since the computational complexity is quadratic in the subgroup size. In future research, we would like to develop bounds for Cook's distance that can be computed for subgroups with small coverage as well.

As we have seen in Section 4, the fraction of subgroups that can be pruned is strongly dependent on the complexity of the regression model we fit. We have seen some datasets (Ames Housing and Wine) for which the model complexities are modest, on which we can prune almost 40% and 30% of the subgroups, respectively. On datasets for which the model complexities are mediocre, we can still prune approximately 20%, and on the dataset for which the model complexity is high, we can prune only 1%.

In Section 5 we have discussed some illustrative examples of subgroups found on datasets from different domains. The models fitted on these subgroups are discussed. These examples show the versatility of the problems which EMM with Cook's distance can solve.

We discussed in Section 3.2 how the joint influence of records makes Cook's distance for single observations theoretically unsuitable for use in a setting where multiple observations are removed simultaneously. However, it may very well be that this problem is not that serious on real-life datasets. Hence, in future research, we would like to see whether we can use Cook's distance for single observations as a proxy for Cook's distance for multiple observations, for instance by summing over D_i for all $i \in I$.

Also in future work, we would like to explore whether we can improve pruning for complex models. Often one is not interested in the influence of all model coefficients, and in Section 3.3 we have seen an adaptation of Cook's distance such that it is evaluated on a subset of the coefficients. We also gave a way to modify the bounds accordingly, but this is done in a rather blunt way. We plan to study whether more sophisticated bounds can be derived, with which we can prune more subgroups.

Finally, this paper was motivated by the Giffen behavior example, in which coefficients not only substantially change in magnitude, but additionally change in sign. Such sign changes can be found on other datasets as well, and the subgroups to which such models are fitted are usually among the most striking subgroups we can find. In future work, we would like to develop a quality measure that explicitly seeks for such sign changes.

Acknowledgments

This research is financially supported by the Netherlands Organisation for Scientific Research (NWO) under project number 612.065.822 (Exceptional Model Mining).

7. REFERENCES

[1] T. Aidt and Z. Tzannatos, Unions and Collective Bargaining, The World Bank, 2002.

[2] C. Bingham, Some identities useful in the analysis of residuals from linear regression, technical report no. 300, School of Statistics, University of Minnesota, St. Paul, 1977.

[3] R. D. Cook, Detection of Influential Observation in Linear Regression, Technometrics 19(1), pp. 15–18, 1977.

[4] R. D. Cook, S. Weisberg, Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression, Technometrics 22(4), pp. 495–508, 1980.

[5] R. D. Cook, S. Weisberg, Residuals and Influence in Regression, Chapman & Hall, London, 1982.

[6] M. Costanigro, R. C. Mittelhammer, J. J. McCluskey, Estimating Class-Specific Parametric Models under Class Uncertainty: Local Polynomial Regression Clustering in an Hedonic Analysis of Wine Markets, Journal of Applied Econometrics 24, pp. 1117–1135, 2009.

[7] C. Dougherty, Introduction to Econometrics (4th edition), Oxford University Press, 2011.

[8] W. Duivesteijn, A. Knobbe, A. Feelders, M. van Leeuwen, Subgroup Discovery meets Bayesian networks – an Exceptional Model Mining approach, Proc. ICDM, pp. 158–167, 2010.

[9] J. Friedman, N. Fisher, Bump-Hunting in High-Dimensional Data, Statistics and Computing 9(2), pp. 123–143, 1999.

[10] J. F. Gentleman, M. B. Wilk, Detecting outliers II: Supplementing the direct analysis of residuals, Biometrics 31, pp. 387–410, 1975.

[11] H. Grosskreutz, S. Rüping, On Subgroup Discovery in Numerical Domains, Data Mining and Knowledge Discovery 19(2), pp. 210–226, 2009.

[12] D. C. Hoaglin, R. Welsh, The hat matrix in regression and ANOVA, American Statistician 32, pp. 17–22, 1978.

[13] R. T. Jensen, N. H. Miller, Giffen Behavior and Subsistence Consumption, American Economic Review 98(4), pp. 1553–1577, 2008.

[14] G. G. Judge, R. C. Hill, W. E. Griffiths, H. Lütkepohl, T.-C. Lee, Introduction to the Theory and Practice of Econometrics (2nd ed.), Wiley, 1988.

[15] W. Klösgen, Subgroup Discovery, Handbook of Data Mining and Knowledge Discovery, ch. 16.3, Oxford University Press, New York, 2002.

[16] D. Leman, A. Feelders, A. J. Knobbe, Exceptional Model Mining, Proc. ECML/PKDD (2) 2008, LNCS, volume 5212, pp. 1–16, Springer, Heidelberg.

[17] A. Marshall, Principles of Economics, MacMillan and co., 1895.

[18] L. Rezende, Econometrics of Auctions by Least Squares, Journal of Applied Econometrics 23, pp. 925–948, 2008.

[19] T. Stengos, E. Zacharias, Intertemporal Pricing and Price Discrimination: A Semiparametric Hedonic Analysis of the Personal Computer Market, Journal of Applied Econometrics 21, pp. 371–386, 2006.