Exceptional Model Mining

Duivesteijn, W.

Author: Duivesteijn, Wouter
Title: Exceptional model mining
Issue Date: 2013-09-17

Chapter 7

Different Slopes for Different Folks – Regression Model

In Chapter 2, we have discussed the Giffen effect. This effect concerns circumstances under which the economic law of demand is broken. Normally, all else equal, demand for a product will decrease if its price increases. However, given certain conditions on the different kinds and relative prices of available food sources (cf. Chapter 2), this relation reverses for the poor but not too poor households: for them the demand for a certain product will *increase* if its price increases. The relation between price of and demand for products is captured by a regression model.

Inspired by this example, we consider the Exceptional Model Mining instance with regression as model class: seeking descriptions for which (a subset of) the parameter vector β significantly deviates from the parameter vector estimated on the whole dataset. The targets $\ell_1, \ldots, \ell_m$ are internally supervised: ℓ_m is the output of the regression model, and $\ell_1, \ldots, \ell_{m-1}$ are the input variables. Formally, we learn the model $Y = X\beta + \varepsilon$, where Y is the $N \times 1$ vector[1] of ℓ_m-values from our dataset, and X is the $N \times m$ full rank matrix of which column 1 consists of N times the value 1 (representing the intercept in the regression model) and the other columns contain the ℓ_i-values from our dataset. So, in matrix form, we have

[1] We explicitly give both dimensions of all vectors for two reasons: on the one hand, to clearly indicate whether the vector comes in row or column form; on the other hand, to facilitate checking that the dimensions match in subsequent matrix products.

$$Y = \begin{pmatrix} \ell_m^1 \\ \ell_m^2 \\ \vdots \\ \ell_m^N \end{pmatrix} \qquad X = \begin{pmatrix} 1 & \ell_1^1 & \ell_2^1 & \cdots & \ell_{m-1}^1 \\ 1 & \ell_1^2 & \ell_2^2 & \cdots & \ell_{m-1}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \ell_1^N & \ell_2^N & \cdots & \ell_{m-1}^N \end{pmatrix} \qquad Y = X\beta + \varepsilon$$

Here, β is the $m \times 1$ vector of the unknown regression parameters, and ε is the $N \times 1$ vector of randomly distributed errors such that $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \mathrm{diag}(\sigma^2 \mathbf{I})$ (where $\mathbf{I}$ 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 [41] to the estimate

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

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

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

We will denote a part of this equation by V, i.e.

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

This matrix was dubbed the *hat matrix* by John W. Tukey, since $\hat{Y} = VY$, i.e. the hat matrix transforms Y into $\hat{Y}$ [51].

7.1 Quality Measure φ_{Cook}

In order to define a proper quality measure for comparing estimated parameter vectors, we need to take into account 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

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

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. Cook originally introduced his distance [13] in 1977 for determining the contribution of single records to $\hat{\beta}$. He states that according to normal theory [42], the $(1-\alpha) \times 100\%$ confidence ellipsoid for the unknown vector, β, is given by the set of all vectors β^* satisfying

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

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

where

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

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

We can exploit the confidence ellipsoid and F-distribution to define a quality measure suitable for the current EMM instance, with some desirable properties. On the one hand, it respects the (co-)variances present in the data, as discussed at the beginning of this section. On the other hand, it comes with theoretical upper bounds that can be utilized to prune the search space, as we will discuss in Section 7.3. To arrive at the definition of the quality measure, however, we first need to determine the degree of influence of single records on the parameter vector. Then we will discuss generalizing this to the influence of deleting multiple records simultaneously. After that, we can give the definition.

Suppose we want to know how a single 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

$$\Delta_i = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)^\top X^\top X \left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{ms^2}$$

Suppose for example that for a certain record r^i we find that $\Delta_i \approx F(m, N - m, 0.5)$. Then removing r^i moves the least squares estimate to the edge of the 50% confidence region for β based on $\hat{\beta}$.

Cook and Weisberg extended Cook's distance to the case where multiple records are deleted simultaneously [14]. Let I be a vector of indices that specify the h records to be deleted. From now on, we let the subscript (I) denote "with the h cases indexed by I deleted", while the subscript I without parentheses denotes "with only the h cases indexed by I remaining". The only notation that deviates from this rule of thumb is the definition of Cook's distance for multiple observations, which becomes

$$\Delta_I = \frac{\left(\hat{\beta}_{(I)} - \hat{\beta}\right)^\top X^\top X \left(\hat{\beta}_{(I)} - \hat{\beta}\right)}{ms^2} \tag{7.1}$$

Its geometric interpretation is identical to the geometrical interpretation of Δ_i. Any subset that has a large joint influence on the estimation of β will result in a large Δ_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.

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 [15] introduce the zero/one-matrix Z, with dimensions $m' \times m$, where m'

is the number of elements of $\hat{\beta}$ that we are interested in (hence $m' \leq m$). 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 m' elements of $\hat{\beta}$, Z will start from the left with $m - m'$ columns containing all zeroes, followed by a $m' \times m'$ identity matrix ($Z = (\mathbf{0}, \mathbf{I}_{m'})$).

When using this transformation, Cook's distance (Equation (7.1)) becomes

$$\Delta_I^{\psi} = \frac{\left(\hat{\beta}_{(I)} - \hat{\beta}\right)^\top Z^\top \left(Z \left(X^\top X\right)^{-1} Z^\top\right)^{-1} Z \left(\hat{\beta}_{(I)} - \hat{\beta}\right)}{m' s^2}$$

Since Cook's distance is invariant to changes in scale of the variables involved [13], it would make an excellent quality measure for use in EMM

Definition (φ_{Cook}). Let D be a description. Its *quality according to Cook's distance* is given by

$$\varphi_{\text{Cook}}(D) = \Delta_I^{\psi}, \text{ where } I = \left\{ i \mid r^i \in \Omega, D\left(a_1^i, \ldots, a_k^i\right) = 0 \right\}$$

The quality of a description according to Cook's distance is the distance bridged when the records *not covered by the description* are simultaneously *discarded*. Hence, Cook's distance is computed for the case where the records covered by the description D are *retained*.

7.2 Experiments

7.2.1 Datasets

The *Giffen Behavior* dataset was used for a study that provided the first real-world evidence of Giffen behavior, i.e. an upward sloping demand curve [77]. As common sense suggests, the demand for a product will normally decrease as its price increases. According to economic textbooks, there are circumstances however, for which the demand curve should slope upward. 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 [53] was collected in a field study in different counties in the Chinese province Hunan, where rice is the staple food. The price changes were brought about by subsidizing the purchase of rice. Each household was randomly assigned to either a control group, or one of three treatment groups. Households in the treatment groups were given vouchers worth ¥0.10, ¥0.20, or ¥0.30, redeemable at selected vendors for a reduction off the price of each *jin* (1 *jin* equals 500 grams) of rice. The average price of rice in Hunan is ¥1.20 per *jin*, so the vouchers represented substantial price changes. The programme provided vouchers for a time period of five months, and subsidized for each person an amount of rice, equal to roughly twice the average per capita consumption.

Data were gathered on three points in time: before the voucher programme started, while the voucher programme was running, and after the voucher programme had ended. Hence, for each household, two changes are observed: the change between the first and second period ($t = 2$), capturing the effect of giving the subsidy; and the change between the second and third period ($t = 3$), capturing the effect of removing the subsidy. The global model estimated in [53] is

$$\%\Delta \mathit{staple}_{i,t} = \beta_0 + \beta_1 \%\Delta p_{i,t} + \sum \beta_2 \%\Delta Z_{i,t} + \sum \beta_3 \mathit{County} \times \mathit{Time}_{i,t} + \Delta\varepsilon_{i,t}$$

where $\%\Delta \mathit{staple}_{i,t}$ 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 $t = 3$), $\%\Delta Z_{i,t}$ is a vector of percent changes in other control variables including income and household size, and *County* × *Time* denotes a set of dummy variables included to control for any county-level factors that change over time. For further details about the design of the study and the estimation strategy, we refer to [53].

The *Ames Housing* 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

$$\mathit{Price} = \beta_0 + \beta_1 \times \mathit{Lot\ Area} + \beta_2 \times \mathit{Quality}$$

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.

The *Auction* dataset was analyzed in [93]. It concerns eBay auctions of Apple iPod mini players from June 27 to July 18, 2006. The goal is to model the final price reached in the auction in terms of auction, seller, and product characteristics. The global model is

$$\begin{aligned} \mathit{Price} &= \beta_0 + \beta_1 \times \mathit{Nbid} + \beta_2 \times \mathit{PositiveFeedback} + \beta_3 \times \mathit{Time} \\ &\quad + \beta_4 \times \mathit{FeedbackScore} + \beta_5 \times \mathit{Memory} + \beta_6 \times \mathit{ResPrice} \end{aligned}$$

where *Price* is the final price of the auction in US dollars, *Nbid* is the number of distinct people who bid in the auction, *PositiveFeedback* is the seller's positive feedback percentage (the coefficient is nonzero from the fourth decimal place), *Time* is the time of he final bid expressed in seconds after Dec. 31 1969, 22:00:00 PDT (the coefficient is nonzero from the fifth decimal place), *FeedbackScore* is the seller's feedback score, *Memory* is the reported memory of the iPod in gigabytes, and *ResPrice* is the auction reservation price in US dollars.

The *EAEF* dataset was extracted from the National Longitudinal Survey of Youth 1979 (NLSY79). It contains information about hourly earnings of men and women, their education, and other information. For more details, see [22, Appendix B]. We fit a model relating years of schooling and years of work experience to earnings in US dollars per hour. The model fitted on the whole dataset is

$$\mathit{Earnings} = \beta_0 + \beta_1 \times \mathit{YrsOfSchool} + \beta_2 \times \mathit{YrsWorkExp}$$

The *Personal Computer* dataset was analyzed in [99]. The data was collected from advertisements in PC Magazine. Each observation consists of the advertised price and features of personal computers. We have learned the following model from the complete dataset

$$\begin{aligned} \mathit{Price} &= \beta_0 + \beta_1 \times \mathit{Spd} + \beta_2 \times \mathit{HD} + \beta_3 \times \mathit{RAM} \\ &\quad + \beta_4 \times \mathit{Scr} + \beta_5 \times \mathit{Ads} + \beta_6 \times \mathit{Trend} \end{aligned}$$

where *Price* is the price in US dollars of a 486 PC, *Spd* is the clock speed in MHz, *HD* is the size of the hard drive in MB, *RAM* is the size of RAM in MB, *Scr* is the size of the screen in inches, *Ads* is the number of 486 price listings in the month the advertisement was placed, and *Trend* is a time trend starting from January of 1993 to November of 1995.

Table 7.1: Statistics concerning the datasets used in the Regression Model experiments. Here, N is the total number of records, k is the number of descriptive attributes , and m is the number of coefficients in the fitted regression model.

Dataset	Domain	N	k	m
Ames Housing	Residential property value	2930	77	3
Auction	eBay auctions	1225	3	7
EAEF	Employment	2714	32	3
Giffen Behavior	Food economics	1254	6	16
Personal Computer	PC pricing	6259	3	7
Wine	Wine pricing	5000	6	4

Finally, the *Wine* dataset was analyzed in [16]. 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 [16]. The global model is

$$\mathit{Price} = \beta_0 + \beta_1 \times \mathit{Cases} + \beta_2 \times \mathit{Score} + \beta_3 \times \mathit{Age}$$

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 have the sign that one would expect based on common sense.

Table 7.1 lists some elementary properties of these datasets.

7.2.2 Experimental Results

Giffen Behavior Data

The global model estimated on the *Giffen Behavior* dataset is

$$\%\Delta \mathit{staple}_{i,t} = \beta_0 + \beta_1 \%\Delta p_{i,t} + \sum \beta_2 \%\Delta Z_{i,t} + \sum \beta_3 \mathit{County} \times \mathit{Time}_{i,t} + \Delta\varepsilon_{i,t}$$

The coefficient of primary interest is β_1. If $\beta_1 > 0$ 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_1.

The authors of [53] 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 [53] tried different manually selected thresholds on *ISCS*; for example, for the subgroup of households with $ISCS > 0.8$, indeed it is observed that $\widehat{\beta}_1 = -0.585$ (no Giffen behavior) whereas for $ISCS \leq 0.8$ they find $\widehat{\beta}_1 = 0.466$ (Giffen behavior).

We analyzed this dataset with *ISCS* as one of the descriptive attributes. The best description we found was D_{14} : $ISCS \geq 0.87$ with $\widehat{\beta}_1 = -0.96$ (against $\widehat{\beta}_1 = 0.22$ for the complete dataset). The coverage of this description is $|G_{14}| = 106$ (3.9%). 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 4th-ranked description D_{15} : *Income per Capita* ≤ 64.67, with a slope of -0.41, and the 6th-ranked description D_{16} : *Income per Capita* ≥ 803.75, with a slope of 0.79 (strong Giffen behavior).

Ames Housing Data

The global model for the *Ames Housing* dataset is

$$Price = -108225.05 + 1.93 \times Lot\ Area + 44201.87 \times Quality$$

By far the most deviating description we find is D_{17}, where the building type is a *'townhouse inside unit'*. For D_{17}, the learned model is

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

The coverage of this description is $|G_{17}| = 101$ (3.4%). 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.

EAEF Data

The global model fitted on the *EAEF* dataset is

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

The 4^{th} ranked description we found was D_{18} : $COLLBARG = 1$, meaning that the pay was set by collective bargaining. The learned model for this description with coverage $|\mathrm{G}_{18}| = 533$ (19.6%) is

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

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].

Personal Computer Data

The global model for the *Personal Computer* dataset is

$$\begin{aligned} Price = &- 246.68 + 8.89 \times Spd + 0.71 \times HD + 47.39 \times RAM \\ &+ 126.70 \times Scr + 0.97 \times Ads - 47.08 \times Trend \end{aligned}$$

By far the most important attribute for defining descriptions was whether or not the company was a *premium firm* (IBM or COMPAQ). The most deviating description was D_{19}, the *non-premium firms*, with model

$$\begin{aligned} Price = & -2130.21 + 13.15 \times Spd + 2.31 \times HD + 22.20 \times RAM \\ & + 252.80 \times Scr + 0.79 \times Ads - 46.45 \times Trend \end{aligned}$$

The coverage of this description is $|G_{19}| = 612$ (9.8%). We get the clearest picture when we contrast this with the regression model fitted to the *premium firms*, which is

$$\begin{aligned} Price = & \ 165.69 + 8.50 \times Spd + 0.67 \times HD + 53.66 \times RAM \\ & + 99.96 \times Scr + 0.65 \times Ads - 47.87 \times Trend \end{aligned}$$

The coverage of this description is $|G_{19}^{C}| = 5647$ (90.2%). We find mostly reasonable behavior in these subgroups: the price of computers from premium firms is based on a far higher intercept, since the premium brand name ensures a vast price upkeep. Consequently, other factors have a substantially smaller impact on the price than for computers from non-premium firms. Oddly, the size of RAM memory does matter more strongly for premium brands than for non-premium brands.

Wine Data

On the *Wine* dataset, the global model is

$$Price = -186.61 - 0.0002 \times Cases + 2.35 \times Score + 5.51 \times Age$$

The most deviating description is D_{20} : *Variety* = *'Non-varietal'* (alternatives are *'Pinot noir'*, *'Cabernet'*, *'Merlot'*, *'Zinfandel'* and *'Syrah'*). The regression model for D_{20} is

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

'Non-varietal' 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: their buyers are more critical.

7.3 Pruning with Bounds for Cook's Distance

Cook's distance is a theoretically well-founded quality measure for mining descriptions for which the slope vector deviates. The bad news is that its computation involves the computation of $\hat{\beta}^G$, which implies that we need to invert a matrix for each candidate description. This is computationally very expensive. Fortunately, some upper bounds have been derived for Cook's distance, which we can use to discard some candidates without having to invert a matrix.

The upper bounds for Cook's distance are derived [15, p. 136] by rewriting the numerator of the right hand side of Equation (7.1) 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 $\lambda_1, \ldots, \lambda_h$, and can assume without loss of generality that $0 \leq \lambda_1 \leq \ldots \leq \lambda_h \leq 1$. Notice that if the last inequality is not strict, i.e. $\lambda_h = 1$, 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 $\operatorname{tr}(V_I)$ here, but notes that this is only a good approximation under the condition that $\operatorname{tr}(V_I) < 1$. Assuming that this condition holds, we can bound D_I by

$$D_I \leq \frac{\operatorname{tr}(V_I)}{(1 - \operatorname{tr}(V_I))^2} \cdot \frac{\sum_{i \in I} e_i^2}{ms^2} \tag{7.2}$$

Unfortunately, this bound is potentially different for each I. Cook also gives bounds that hold for all subsets I of a fixed size h. When we fix h 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 (7.2) into

$$D_I \leq \frac{\operatorname{tr}(V_I)}{(1 - \operatorname{tr}(V_I))^2} \cdot \frac{R^2}{ms^2} \tag{7.3}$$

or we could use $T = \max_I (\operatorname{tr}(V_I))$, which turns Equation (7.2) into

$$D_I \leq \frac{T}{(1 - T)^2} \cdot \frac{\sum_{i \in I} e_i^2}{ms^2} \tag{7.4}$$

Both estimates can be combined to turn Equation (7.2) into

$$D_I \leq \frac{T}{(1-T)^2} \cdot \frac{R^2}{ms^2} \tag{7.5}$$

Rather obviously, there are relations between the bounds: bound (7.2) is stricter than both bound (7.3) and bound (7.4), and those are both stricter than bound (7.5).

Whenever one has the possibility to enumerate all candidate descriptions for mining with Cook's distance, the bounds (7.2)–(7.5) can be used for pruning. In combination with the beam search strategy for top-q EMM, we propose to do this in the following way.

Per search level, we enumerate all candidate descriptions in descending order according to bound (7.5). Then we consider the subgroups in this order. For each description, we compute the bounds in order of decreasing ease of computation, i.e. first bound (7.5), then bound (7.4), then bound (7.3), and finally bound (7.2). We check whether any of these bounds has a value that is lower than Cook's distance for the q^{th} best evaluated description so far. If so, we know that Cook's distance for this new description can not enter the top-q, since the bound is an upper bound for Cook's distance. Hence we can skip computing Cook's distance for this description, 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 description.

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

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

The results can be found in Figure 7.1. The figure depicts the subset size on the x-axis (linear scale), and the values of the bounds on the y-axis (logarithmic scale).

The *EAEF* dataset has 2714 records, so when a subset 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 subset

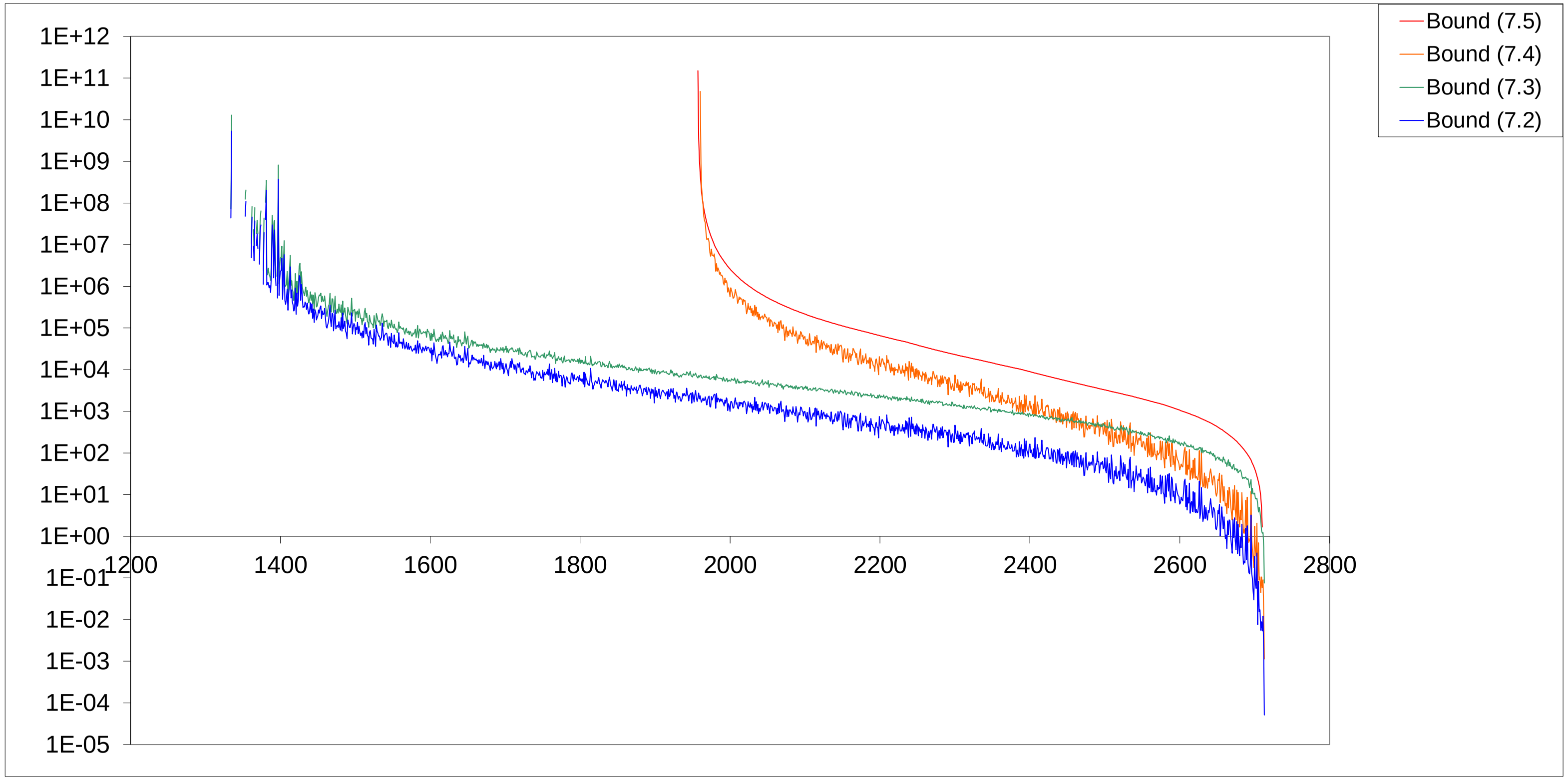


Figure 7.1: Bound values (logarithmic scale) for random subsets of different sizes on the *EAEF* dataset. Fitted model: $Earnings = \beta_0 + \beta_1 \times YrsOfSchool$.

size 0. This is caused by limitations in the approximations used in the bounds. As mentioned before, 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 (7.4) and (7.5), 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 subsets having less than 1960 records, bounds (7.4) and (7.5) cannot be computed. For subsets having less than roughly 1250 records, this also holds for bounds (7.2) and (7.3). 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 (7.2) and (7.3) can only be computed when the subset contains at least 50% of the records, and bounds (7.4) and (7.5) only when the subset 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 subsets 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 subsets, and the computation of the hat matrix is quadratic in the subset size. Hence whenever we can prune a subset, it always takes a relatively expensive regression computation out of the total runtime.

7.3.1 Empirical bound evaluation

To empirically see how the bounds function, we performed a depth-1 EMM run on each dataset, with the goal to find the top-1 description. When numeric attributes were used to generate candidate descriptions, we split them into 12 equal-sized bins. We discarded any description covering fewer 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 description to be pruned.

The results can be found in Table 7.2a. This table features the datasets, dataset characteristics, number of times every bound is computed, number of descriptions pruned with every bound, fraction of candidate descriptions for which at least one bound was computable, and fraction of candidate

Table 7.2: Pruning results for depth-1 EMM runs. Here, N is the total number of records, $\mathcal{C}$ is the set of candidate descriptions considered, and m is the number of coefficients in the fitted regression model. The set "bounded $\mathcal{C}$" consists of the candidate descriptions for which at least one bound could be computed, and the set "pruned $\mathcal{C}$" consists of the candidate descriptions that were pruned using one of the bounds.

(a) Results when looking for the top-1 description.

Dataset	N	$\lvert\mathcal{C}\rvert$	m	Bounds computed				Descriptions pruned				$\frac{\lvert\text{bounded } \mathcal{C}\rvert}{\lvert C\rvert}$	$\frac{\lvert\text{pruned } \mathcal{C}\rvert}{\lvert\mathcal{C}\rvert}$
				(7.5)	(7.4)	(7.3)	(7.2)	(7.5)	(7.4)	(7.3)	(7.2)		
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

(b) Results when looking for the top-50 descriptions.

Dataset	N	$\lvert\mathcal{C}\rvert$	m	Bounds computed				Descriptions pruned				$\frac{\lvert\text{bounded } \mathcal{C}\rvert}{\lvert C\rvert}$	$\frac{\lvert\text{pruned } \mathcal{C}\rvert}{\lvert\mathcal{C}\rvert}$
				(7.5)	(7.4)	(7.3)	(7.2)	(7.5)	(7.4)	(7.3)	(7.2)		
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

descriptions that were pruned. Notice that there is a strong dependency between the "Bound computed" and "Descriptions pruned" columns: in the *Ames Housing* dataset we can compute bound (7.5) for 196 descriptions, of which we can prune 155, so only 41 subgroups remain for which we compute bound (7.4). However, the number of descriptions for which we compute bound (7.3) is larger, since the condition under which this bound is computable is less strict than the condition for bound (7.4) and (7.5). Of the 228 descriptions for which we compute bound (7.3) we can prune 191, leaving 37 descriptions for which we compute bound (7.2).

As we indicated earlier, the fraction of descriptions 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 $3 \leq m \leq 4$ we can compute bounds for over 40% of the descriptions, in the datasets for which $m = 7$ we can compute bounds for $33 - 35\%$ of the descriptions, and in the dataset for which $m = 16$ we can compute bounds for just 1% of the descriptions. This dependency becomes somewhat less direct when we look at the percentage of descriptions 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 description, we replicate these experiments with the goal to find the top-50 descriptions. Since we need to have considered at least 50 descriptions before we can make sure others 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 7.2b encompasses the results of these experiments on merely the *Ames Housing* and *EAEF* dataset. Notice that the fraction of descriptions we can prune on the *Ames Housing* dataset has only decreased slightly, while the fraction of descriptions we can prune on the *EAEF* dataset is cut in half.

We repeat all these experiments with depth-2 EMM runs with beam width $w = 10$. We find that in these experiments, we can barely compute bounds for any level-2 descriptions, let alone prune them. This is caused by the fact that level-2 descriptions are refinements of well-scoring level-1 descriptions, which usually cover relatively few records. Such descriptions scarcely ever cover more than 50% of the records, hence their refinements also scarcely ever do so. Fortunately, that also means that the regression computations for these level-2 descriptions is relatively cheap.

7.4 Alternatives

We can define a simpler, statistically founded quality measure when we restrict ourselves to a simpler regression model, allowing only one input ($y = \ell_2$) and one output variable ($x = \ell_1$) in the regression, i.e.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{7.6}$$

Consider model (7.6) learned from a subgroup G and its complement G^C. Of course, there is a choice of distance measures between the fitted models. We propose to look at the difference in the slope β_1 between the two models, because this parameter is usually of primary interest when fitting a regression model: it indicates the change in the expected value of y, when x increases with one unit. Another possibility would be to look at the intercept β_0, if it has a sensible interpretation in the application concerned. As with the correlation model, we use significance testing to measure the distance between the fitted models. Let β_1^G be the slope for the regression function of G and $\beta_1^{G^C}$ the slope for the regression function of G^C. The hypothesis to be tested is

$$H_0 : \beta_1^G = \beta_1^{G^C} \quad \text{against} \quad H_1 : \beta_1^G \neq \beta_1^{G^C}$$

We use the least squares estimate $\hat{\beta}_1$ for the slope β_1, and unbiased estimator s^2 for the variance of $\hat{\beta}_1$, i.e.

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad s^2 = \frac{\sum \hat{\varepsilon}_i^2}{(\xi - 2) \sum (x_i - \bar{x})^2}$$

where $\hat{\varepsilon}_i$ is the regression residual for individual i, and ξ is the sample size. Finally, we define our test statistic t'. Although it does not have a t distribution, its distribution can be approximated quite well by one, with degrees of freedom given at the top of the next page (cf. [81])

$$t' = \frac{\hat{\beta}_1^G - \hat{\beta}_1^{G^C}}{\sqrt{s^{G^2} + s^{G^{C2}}}} \qquad\qquad df = \frac{\left(s^{G^2} + s^{G^{C2}}\right)^2}{\frac{s^{G^4}}{n-2} + \frac{s^{G^{C4}}}{n^C-2}}$$

The approximation is accurate when $n + n^C \geq 40$ (cf. [81]), so unless we analyze a very small dataset we should be confident to base p-value computation on it. Our quality measure φ_{ssd} (acronym for Significance of Slope Difference) is one minus this p-value.

Running EMM on the *Windsor Housing* dataset (cf. Table 4.1) using φ_{ssd} as quality measure, we find as first-ranked description D_{21} the 226 houses (41.3% of the dataset) that have a *driveway*, *no basement* and *at most one bathroom*

$$D_{21} : \textit{drive} = 1 \wedge \textit{basement} = 0 \wedge \textit{nbath} \leq 1$$

From the subgroup G_{21} and its complement G_{21}^C (320 houses, 58.7%) we learn the following two regression functions, respectively

$$\begin{aligned} G_{21} &: \; y = 41568 + 3.31 \cdot x \\ G_{21}^C &: \; y = 30723 + 8.45 \cdot x \end{aligned}$$

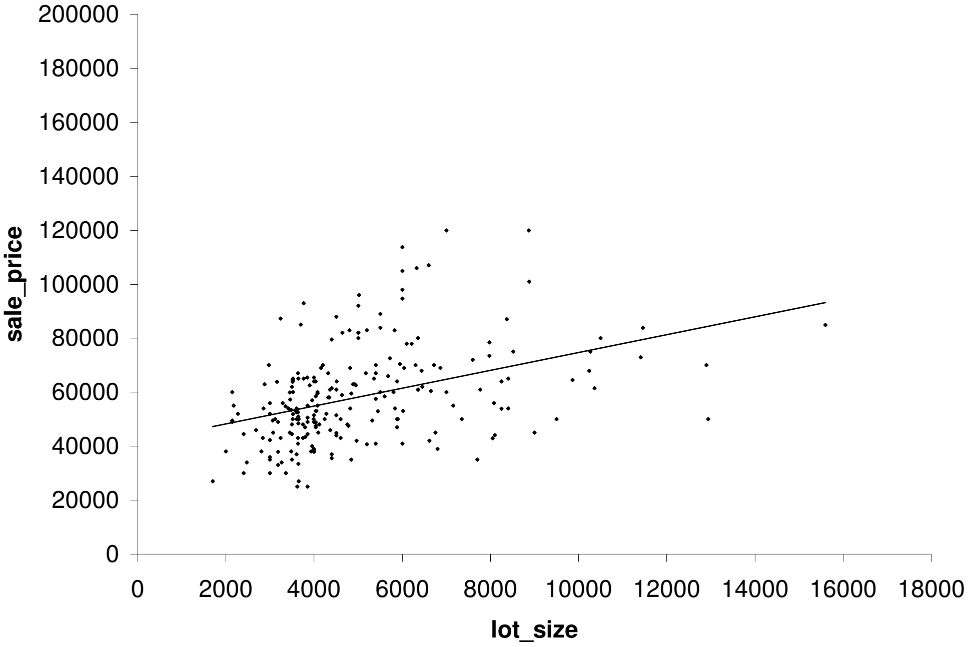
The description quality is $\varphi_{\text{ssd}}(D_{21}) > 0.9999$, meaning that the p-value of the test

$$H_0 : \beta_1^{G_2} = \beta_1^{G_2^C} \quad \text{against} \quad H_1 : \beta_1^{G_2} \neq \beta_1^{G_2^C}$$

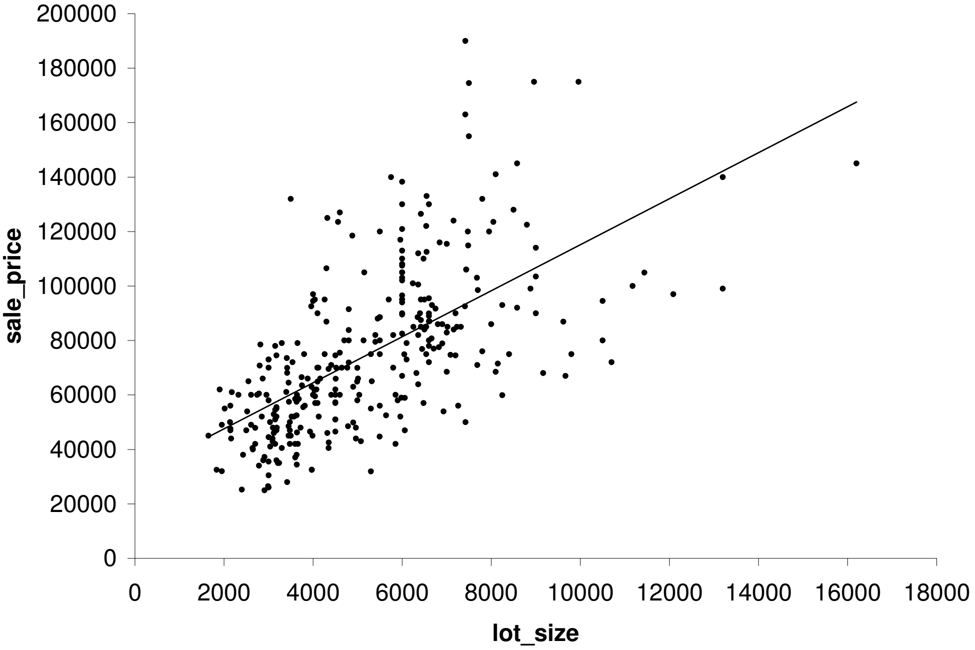
is virtually zero. There are descriptions with a larger difference in slope, but the reported description scores higher because its coverage is quite big. Figure 7.2 shows the scatter plots of *lot_size* and *sales_price* for the description and its complement.

7.5 Conclusions

In this chapter, we propose to use Cook's distance in an Exceptional Model Mining setting. This allows us to find descriptions, for which a regression model fitted on the targets is substantially different from that model for the whole dataset. The use of Cook's distance has two benefits.



(a) $G_{21}, y = 41568 + 3.31 \cdot x$.



(b) $G_{21}^{C}, y = 30723 + 8.45 \cdot x$.

Figure 7.2: *Windsor Housing* - φ_{ssd}: Scatter plot of *lot_size* and *sales_price* for the subgroup corresponding to D_{21} : *drive* $= 1 \wedge$ *basement* $= 0 \wedge$ *nbath* ≤ 1 and its complement.

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 $\widehat{\beta}$ 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 description covering at least 50% of the records. On the one hand, this means that we can compute the bounds for relatively few descriptions, but on the other hand, whenever we can prune a description, we always prune a relatively expensive regression computation. In future research, we would like to develop bounds for Cook's distance that can be computed for descriptions with small coverage as well.

As we have seen in Section 7.3, the fraction of descriptions 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 descriptions, respectively. On datasets whose 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 7.2.2 we have discussed some illustrative examples of descriptions found on datasets from different domains. The models fitted on these descriptions are discussed. These examples show the versatility of the problems which EMM with Cook's distance can solve.

Theoretically, 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 at the end of Section 7.1 we have seen an adaptation of Cook's distance such that it is evaluated on a subset of the coefficients. Modifying the bounds accordingly is done in a rather blunt way. We plan to study whether more sophisticated bounds can be derived, with which we can prune more descriptions.

Finally, this chapter 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 descriptions to which such models are fitted are usually among the most striking deviations we can find. In future work, we would like to develop a quality measure that explicitly seeks for such sign changes.