STAT 383C: Statistical Modeling I **Fall 2015**

Lecture 15 — October 15

Lecturer: Purnamrita Sarkar *Scribe: Prateek Srivastava*

Disclaimer: These scribe notes have been slightly proofread and may have typos etc.

Note: The latex template was borrowed from EECS, U.C. Berkeley.

15.1 Influence of Data Points

In order to understand the influence that a data point has on regression, we first define the terms of leverage and discrepancy.
Leverage: If a data point has an unusually high or low value of X from its mean, then it is likely to have higher leverage on the regression curve.
Discrepancy: If y_i is unusual for a given x_i, then the observation is said to have a high discrepancy.

From these two definitions, we can now define the concept of influence of a data point as Influence = Leverage X Discrepancy.

Influential data points with high leverage or discrepancy may result in poor predicted values from regression y. In order to identify such points, therefore, Cook's distance (D_i) is used as a measure to evaluate the influence of a data point.

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_j^{(-i)}\right)^2}{p\text{MSE}} \tag{15.1}$$

In the above expression, $\hat{y}_j$ is the predicted value for observation j from the full regression model, $\hat{y}_j^{(-i)}$ is the predicted value for observation j from the regression model with omitted observation i, p is the number of fitted parameters in the model and MSE is the mean squared error of the regression.

15.2 Pros and Cons of M-estimators

PROS

1. Almost as efficient as ordinary least squares under the Gaussian model.
2. Computationally efficient.
3. More robust.

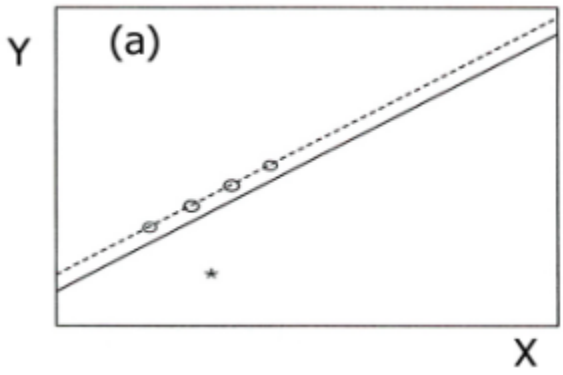

Figure 15.1: **Outlier without influence.** Although its Y value is unusual given its X value, it has little influence on the regression line because it is in the middle of the X- range.

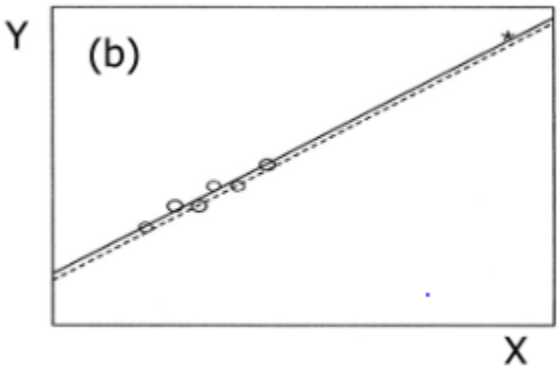

Figure 15.2: **High leverage** because it has a high value of X. However, because its value of Y puts it in line with the general pattern of the data it has no influence.

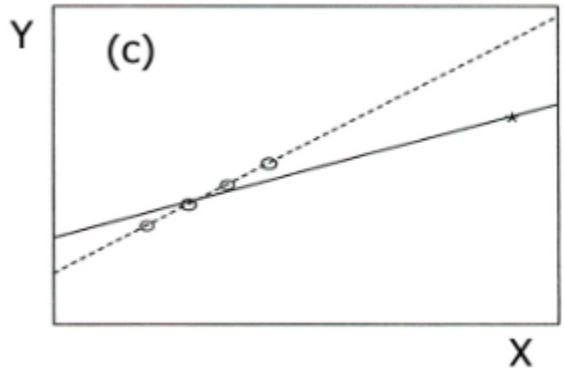

Figure 15.3: **Combination of discrepancy (unusual Y value) and leverage (unusual X value) results in strong influence.** When this case is deleted both the slope and intercept change dramatically.

Figures and Captions taken from http://polisci.msu.edu/jacoby/icpsr/regress3/lectures/w
CONS

1. M estimators have a low breakdown point (≈ 0) and therefore, may suffer when we have a high leverage, high discrepancy point.

The alternatives to using M-estimators for robust regression are bounded influence regression techniques like Least Trimmed Squares or Least Median of Squares. The breakdown point for such methods can be as high as 50%. However, this comes at the cost of throwing away the efficiency. In addition, these methods are also computationally expensive.

Thus far, we have studied prediction techniques for which both Xs and Ys were given to us. In linear regression, for example, we had both continuous Xs and continuous Ys. In logistic regression and Naive Bayes, the Ys are binary and the Xs are continuous. These techniques are examples of supervised learning techniques.

15.3 K-means clustering

Given a fixed number of clusters K and set of observations $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$, the K-means clustering problem is the problem of partitioning the n observations into K clusters such that the within-cluster sum of squares (WCSS) distance is minimized. Mathematically, the optimization problem can be formulated as follows:

$$\min_{\mathbf{S}} \sum_{k=1}^{K} \sum_{i \in S_k} \|x_i - \mu_k\|^2$$

In the above formulation, $\mathbf{S} = \{S_1, ..., S_K\}$ denotes the set of K clusters and μ_k is the mean of points in cluster S_k .

15.3.1 Algorithm

1. Initialize by randomly assigning K cluster centers: $\mu_1^{(0)}, \mu_2^{(0)}, ..., \mu_{\mathbf{K}}{}^{(0)}$.

2. Assign each data point to its closest cluster center to obtain $S_i^{(t)}$ as follows:

$$S_k^{(t)} = \{x_i : \|x_i - \mu_k^{(t)}\|^2 \leq \|x_i - \mu_j^{(t)}\|^2 \ \forall j, 1 \leq j \leq K\} \tag{15.2}$$

3. Recalculate the mean $\mu_k^{(t+1)}$ as follows:

$$\mu_k^{(t+1)} = \frac{1}{|S_k^{(t)}|} \sum_{i \in S_k(t)} x_i \tag{15.3}$$

4. Repeat steps 2 and 3 until convergence in the means is obtained.