Building predictive models for big data using publicly available data

2nd Texas A&M Big Data Workshop 2016

Professor Simon Sheather  
email: sheather@stat.tamu.edu
Outline

• Predictive models for two large public data sets -
  1. The NYC Taxi and Limousine Commission trip record data for every trip from 1/1/2009 until 6/30/2015.
  2. Airline Origin and Destination Survey (DB1B), which consists of a 10% sample of all airline ticket sold each quarter dating back to 1993.

• The illusion of apparently very high precision
> 1.1 billion individual taxi trips:

Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

Data Profile: Airline Origin and Destination Survey (DB1B)

Overview

The Airline Origin and Destination Survey (DB1B) is a 10% sample of airline tickets from reporting carriers collected by the Office of Airline Information of the Bureau of Transportation Statistics. Data includes origin, destination and other itinerary details of passengers transported. This database is used to determine air traffic patterns, air carrier market shares and passenger flows.

Robust regression estimates

PROC ROBUSTREG in SAS 9.4 (with each method based on the default settings)
1. M-estimate
2. Least trimmed squares (LTS) estimate
3. LTS FWLS estimate
4. S-estimate
5. MM-estimate

Plus a robust rank-based estimate obtained by a referee using the R software package
Robust regression estimates

2.1 M-estimates

An M-estimate \( \hat{\theta}_M \) of \( \theta \) (Huber, 1973) minimizes the following sum

\[
Q_M(\theta) = \sum_{i=1}^{n} \rho \left( \frac{r_i}{\sigma} \right)
\]

2.2 LTS estimate

The least trimmed squares (LTS) estimate \( \hat{\theta}_{LTS} \) of \( \theta \) (Rousseeuw, 1984) minimizes the following sum

\[
Q_{LTS}(\theta) = \sum_{i=1}^{h} r_{(i)}^2
\]

where \( r_{(1)}^2 \leq r_{(2)}^2 \leq \cdots \leq r_{(n)}^2 \) are the ordered squared residuals and \( h \) is defined in the range \( \frac{n}{2} + 1 \leq h \leq \frac{3n+p+1}{2} \).

2.3 S estimate

The S estimate \( \hat{\theta}_S \) of \( \theta \) (Rousseeuw and Yohai, 1984) minimizes the dispersion \( S(\theta) \) where \( S(\theta) \) is the solution of

\[
\frac{1}{n-p} \sum_{i=1}^{n} \chi \left( \frac{y_i - x_i^T \theta}{S} \right) = \beta
\]

where \( \beta = \int \chi(s) \, d\Phi(s) \) so that \( \hat{\theta}_S \) and \( S(\hat{\theta}_S) \) are asymptotically consistent estimates of \( \theta \) and \( \sigma \) for the Gaussian regression model. The breakdown value of the S estimate is equal to \( \beta / \sup_{s} \chi(s) \).

2.4 MM estimate

The MM estimate \( \hat{\theta}_{MM} \) of \( \theta \) (Yohai, 1987) is based on a combination of the use of high breakdown estimation and efficient estimation procedures. MM estimate with an LTS initial estimate.
NYC Taxi Trip Data

In this study we shall focus on data for taxi trips taken on a randomly selected day in January, 2013, namely Tuesday January 15, 2013. In particular, we shall consider $n = 49,800$ taxi trips with the following characteristics:

• rate_code = 1, which corresponds to the standard city rate
• rounded_trip_distance < 3 miles, where the rounding was down to the nearest 1/5 mile
• average_trip_speed ≥ 25 miles per hour

For rate code 1, the initial charge is $2.50 plus 50 cents per 1/5 mile or 50 cents per 60 seconds in slow traffic or when the vehicle is stopped. “slow traffic” is defined to be travelling under 12 miles an hour.
The median(fare_amount) is a linear function of rounded_trip_distance. In particular,

\[
\text{median(fare_amount)} = 2.50 + 2.50 \times \text{rounded_trip_distance} \quad (1)
\]

This is to be expected since the fare structure is such that the initial charge is $2.50 plus 50 cents per 1/5 mile.
NYC Taxi Trip Data

Conclusions:
1. Only the M-estimates and the R-estimates are equal to the values of the intercept and the slope in (1), namely, $2.50.
2. The confidence intervals are very narrow implying high precision of the point estimates.
Air fare data

• The DB1BTicket file contains data on 3,588,928 flight itineraries involving 7,021,913 passengers. We shall focus on $n=78,905$ single passenger nonstop round trip flight itineraries on Southwest Airlines in the contiguous domestic market.

• We seek to build a model for ItinFare, the itinerary fare per person from MilesFlown, the miles flown according to the flight itinerary.
Denote ItinFare by $Y$ and MilesFlown by $x$. We considered regression spline models of the form

$$Y = \beta_0 + \beta_1(1500 - x)_- + \beta_2(x - 1500)_+$$  \hspace{1cm} (2)

where

$$(1500 - x)_- = \begin{cases} x - 1500, & x < 1500 \\ 0, & x \geq 1500 \end{cases}$$

and

$$(x - 1500)_+ = \begin{cases} 0, & x < 1500 \\ x - 1500, & x \geq 1500 \end{cases}$$
Conclusions:
1. The estimates of the 2 slope parameters vary widely between methods.
2. The confidence intervals are very narrow implying high precision of the point estimates.
Air fare data

In the analyses presented, no account was taken of the fact that airfares vary across many factors including:

• Time of the day
• Day of the week
• The two airports that the flights are between
• The number of days before the flight during which the ticket was purchased
• How many vacant seats exist on the flight at the time of booking

Thus, it is reasonable to conclude that the regression coefficients in model (2) can be expected to take very different values in different combinations of these factors. For example, compare and contrast the airfare for a ticket that is purchased the day of the flight with very few vacant seats at the busiest time of the day between two airports between which there is little competition between carriers the airfare for a ticket that is purchased long before the day of the flight with very many vacant seats at the least busy time of the day between two airports between which there is a great deal of competition between carriers. There is likely to be a very substantial difference between these two airfares. In addition, there is likely to be strong dependence between the airfare of tickets purchased with similar combinations of these factors.
The illusion of apparently very high precision

In a recent paper, Cox (2015) finds that

- “So-called big data are likely to have complex structure, in particular implying that estimates of precision obtained by applying standard statistical procedures are likely to be misleading. ... With very large amounts of data, direct use of standard statistical methods ... will tend to produce estimates of apparently very high precision, essentially because of strong explicit or implicit assumptions of at most weak dependence underlying such methods. ... The most serious possibility of misinterpretation arises when the regression coefficient takes very different values in the different base processes.”

In addition, Cox (2015) recommends that

- We ... “consider big data as evolving in a possibly notional time-frame. At various time-points new sources of variability enter” ... and that we ... “represent the main sources of variation in an explicit model and thereby produce both improved estimates and more relevant assessments of precision".
Business Analytics & Statistical Modeling Blog at Texas A&M

Welcome to the Business Analytics & Statistical Modeling Blog hosted by Texas A&M MS Analytics. This blog will highlight and discuss the findings from predictive modeling of topical events in analytics across a wide spectrum of domains.

February 2, 2016
In the NFL, the number of concussions experienced by players differs across divisions within conferences

http://analytics.stat.tamu.edu/basmblog.php