Flexible Statistical Modeling Methods for Big Data

April 21, 2017

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

• Personal history
• “Big data” and predictive models
• Modeling non-zero 12 month gas production
  – Marginal model plots
• Multiple adaptive regression splines (MARS)
  – Modeling NFL fan ratings of games
• Robust regression models of NYC taxi fares and airline ticket prices
  – The illusion of apparently very high precision
• Regression models with time series errors
  – Monthly Chicago Taxi Fare Totals per Medallion
• Transfer function models
  – Modeling CA$ exchange rate as a function of oil price
• Student project examples
  – Predicting weekly US rig count
I was born and educated in Melbourne, Australia
In February 2005, I moved from Sydney, Australia to College Station, Texas
Ancestry

- **Henry Sheather** was born October 22, 1797 in Brede, Sussex, England, and died May 16, 1865 in Redfern, Sydney, New South Wales.
- Immigration depart: 1838, Royal George ex Gravesend, England
- Immigration arrive: March 10, 1839, Sydney.
- Occupation: agricultural laborer (*who could read and write*).
- One of two brothers who came to Australia to work for James Macarthur at Camden Park.
The 4 Sheather Brothers

*Which one am I?*
The 4 Sheather Brothers

Which one am I?

Andrew, Contractor
Simon, Professor
Martin, Banker
Tim, Car Racing
Head of the Department of Statistics at Texas A&M from March 1, 2005 until February 28, 2014
In Fall 2007, MS (Statistics) online began with 20 students

What we offer
- Master of Science in Statistics
- Applied Statistics Certificate
- Individual Courses (see list)

What Sets Us Apart
- All courses 100% online
- Start any semester
- Online office hours
- Wide range of electives

Application Deadlines
- Aug 1 – Fall Start
- Dec 13 – Spring Start
- Apr 30 – Summer Start

Requirements
- Calculus I and Calculus II
- GRE scores (may be waived)
- Statement of Purpose
- Letters of recommendation

Providing a Master's in Statistics Online Since 2007
We have over 200 graduates successfully complete the online program and receive a Texas A&M diploma. Our graduates get the same degree and diploma as the local students. They have even ordered an Aggie ring. Most of them kept working full-time while acquiring their degree.
Texas A&M Statistical Services LP was formed in 2012

Our Services

Texas A&M Statistical Services provides high-quality services in business analytics and the application of statistics to big data problems.

We are dedicated to helping organizations harness data and apply analytics to product design, service improvement, marketing and decision-making.

Our services include:

- Business Assessments for Analytics
- Public Webinars & Workshops
- Customized Corporate Training
- Analytics Support for Mission Critical Projects
- Customer Satisfaction Analysis & Tracking

http://www.tamstatservices.com/
In Fall 2013, MS (Analytics) program in partnership with the Mays Business School

<table>
<thead>
<tr>
<th>ATM</th>
<th>STATISTICS</th>
<th>PART TIME</th>
<th>ANALYTICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
</tr>
</tbody>
</table>
| An analytics degree from Texas A&M, and two years of training with some of the best professionals in the industry will help you know the value of your data. | MODERN SKILL SET
BUSINESS ACUMEN
DEEP ANALYTICAL SKILLS
TECHNICAL EXPERTISE
SOFT SKILLS | Turning Data Into Better Decisions and Business Results |
| Excellent ROI | IN PERSON OR ONLINE
SAME EDUCATION |
| In Partnership with Mays Business School | Instruction/Course Materials
Homework/Exams
Degree |
| Work-Based Project | SAME |
| Flexible Delivery | Masters of Science in Analytics |

www.analytics.stat.tamu.edu
Definitions of “big data”

“There is *no rigorous definition of big data*. Initially the idea was that the *volume of information* had grown so large that the quantity being examined no longer fit into the memory that computers used for processing ....” (page 6)
“N = all” definition of big data

“In many areas, however, a shift is taking place from collecting some data to gathering as much as possible, and if feasible getting everything: \( N = \text{all} \).” (page 26)

“Using all the data need not be an enormous task. Big data is not necessarily big in absolute terms, although often it is.” (page 28)
Two broad types of statistical modeling

- **Explanatory modeling** is the process of building and applying a statistical model that is **interpretable**. In other words, determining which predictors have a meaningful effect on the outcome variable as well as understanding each of these effects.
  - A lender in Texas that uses a model to screen customers has to be able to explain to a potential customer why their loan application was not approved

- **Predictive modeling** is the process of building and applying a statistical model to data in order to **predict** new or **future** observations
  - A credit card company wants to predict in real time whether a credit card transaction is fraudulent or not
The Best Explanatory Models are *Sophistically Simple*

Some years ago, I came upon the phrase used in industry, “Keep It Simple Stupid,” that is KISS and thought about it in relation to scientific model-building. Since some simple models are stupid, I decided to reinterpret KISS to mean “Keep It Sophistically Simple.”

Arnold Zellner, University of Chicago

... it is well known that Einstein advised in connection with theorizing in the natural sciences, “*Make it as simple as possible but no simpler*.”
Predictive Analytics

“Predictive analytics encompasses a variety of techniques from statistics, modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

Predictive analytics is used in actuarial science, marketing, financial services, insurance, telecommunications, retail, travel, healthcare, pharmaceuticals and other fields.”

Source: http://en.wikipedia.org/wiki/Predictive_analytics
Subject Matter Expertise is Important in Model Development
Modeling 12 month non-zero gas production


- The primary modeling goal is to understand which operational variables most impacted well performance, with an initial focus on both proppant and fracture fluid volumes. ... Proppant is a large cost factor in the unconventional drilling process; the optimization of proppant usage will lead to substantial savings.

Source: [http://blumtexas.blogspot.com/](http://blumtexas.blogspot.com/)
Modeling 12 month non-zero gas production

In phase 1 the variable selection step was initiated that implemented a sequential R-square algorithm. The input variables were sequentially selected to explicate the most variation in production data and the results enumerated:

- County (grouped)
- Total Depth
- Y Coordinate (16 bins)
- Frac Fluid (16 bins)
- X Coordinate (16 bins)
- Proppant Volume
- Proppant Volume (16 bins)
- Gross Perforated Interval (16 bins)
- Upper Perforation (16 bins)
- Total Depth (16 bins)
- Frac Fluid
- Lower Perforation (16 bins)

I also added the following predictors:

- Carbonate
- Number of Stages
Modeling 12 month non-zero gas production

- The outcome variable and each of the predictors, apart from Number of Stages, was transformed using a log transformation.
- This reduced skewness and it will allow for estimates of %change effects.
- The only highly correlated predictors are Log[lower perforation_xy] and Log[upper perforation_xy].
Modeling 12 month non-zero gas production

Red – Training data (65%)
Blue – Validation data (35%)

Conclusions:
•
•
Modeling 12 month non-zero gas production – only first order terms

Is this first order model valid?
Modeling non-zero 12 month gas production – only first order terms
Marginal model plot to check model validity

Conclusion:
• First order model for Log[Proppant_LB] is not valid, since the two curves do not match
Modeling non-zero 12 month gas production – first & second order terms
Modeling non-zero 12 month gas production – Marginal model plot

Conclusion:
- Second order model for Log[Proppant_LB] is valid, since the two curves match reasonably well.
Modeling non-zero 12 month gas production – first & second order terms

Recall that the initial focus was on both proppant and fracture fluid volumes. The second order model finds that Log[nzd_12_cum] is maximized for high values of proppant and low values of fracture fluid volumes.
MARS

(Multiple adaptive regression splines)

MARS uses expansions in piecewise linear basis functions of the form $(x - t)_+$ and $(t - x)_+$. The “+” means positive part, so

$$(x - t)_+ = \begin{cases} x - t, & \text{if } x > t, \\ 0, & \text{otherwise}, \end{cases} \quad \text{and} \quad (t - x)_+ = \begin{cases} t - x, & \text{if } x < t, \\ 0, & \text{otherwise}. \end{cases}$$

**FIGURE 9.9.** The basis functions $(x - t)_+$ (solid orange) and $(t - x)_+$ (broken blue) used by MARS.
FIGURE 9.10. Schematic of the MARS forward model-building procedure. On the left are the basis functions currently in the model: initially, this is the constant function. On the right are all candidate basis functions to be considered in building the model. These are pairs of piecewise linear basis functions as in Figure 9.9. ... At each stage we consider all predictors and basis pairs. The basis pair that decreases the residual error the most is added into the current model. Above we illustrate the first three steps of the procedure, with the selected functions shown in red.
2009-2010 NFL Fan Ratings

“For the 2009-2010 NFL season, visitors to the NFL.com website were offered an opportunity to view detailed statistics for each individual game. ... The NFL asks fans for input by “rating” individual games. Fans are simply asked to rate the game on a scale of 0–100, with 0 being Forgettable, and 100 being Memorable by selecting where a needle should be placed on a gauge. No further instructions are offered ... The question we wish to investigate for this article is what determines fan satisfaction with individual NFL games, as measured by each game’s fan rating. ... These ratings were compiled at the end of the season to obtain a complete listing of all games played in the NFL during the 2009-2010 season. ...”

Source: Rodney J. Paul, Yoav Wachsman, and Andrew P. Weinbach entitled “The Role of Uncertainty of Outcome and Scoring in the Determination of Fan Satisfaction in the NFL” which was published in the Journal of Sports Economics in December 2011. We shall refer to this paper as PWW (2011).
## 2009-2010 NFL Fan Ratings

### Table 2. Regression Results—Determinants of Fan Ratings of NFL Games

<table>
<thead>
<tr>
<th>Dependent Variable: Fan Rating</th>
<th>I</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>32.8918*** (11.4755)</td>
<td>30.4463*** (9.5246)</td>
</tr>
<tr>
<td>Margin of victory</td>
<td>-0.2799*** (-4.1641)</td>
<td>-0.2585*** (-3.7835)</td>
</tr>
<tr>
<td>Combined score of both teams</td>
<td>0.5303*** (10.3084)</td>
<td>0.5395*** (10.4621)</td>
</tr>
<tr>
<td>Sum of win percentage</td>
<td>13.5451*** (6.7566)</td>
<td>13.1384*** (6.3733)</td>
</tr>
<tr>
<td>Overtime dummy</td>
<td>9.1509*** (2.4642)</td>
<td>9.6273*** (2.5699)</td>
</tr>
<tr>
<td>October</td>
<td>-1.2760 (-0.6401)</td>
<td></td>
</tr>
<tr>
<td>November</td>
<td>1.4521 (0.7825)</td>
<td></td>
</tr>
<tr>
<td>December/January</td>
<td>1.3357 (0.6928)</td>
<td></td>
</tr>
<tr>
<td>Fox network</td>
<td>-1.1767 (-0.7892)</td>
<td></td>
</tr>
<tr>
<td>NBC network</td>
<td>0.9640 (0.3296)</td>
<td></td>
</tr>
<tr>
<td>ESPN network</td>
<td>1.6126 (0.5525)</td>
<td></td>
</tr>
<tr>
<td>NFL network</td>
<td>-1.9023 (-0.4738)</td>
<td></td>
</tr>
<tr>
<td>LATE (4:05 or 4:15 Sunday Start)</td>
<td>4.1636** (2.6587)</td>
<td></td>
</tr>
<tr>
<td>Division game</td>
<td>2.4079* (1.6824)</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.4682</td>
<td>0.4977</td>
</tr>
</tbody>
</table>

* Represents significance at 10%; ** represents significance at 5%; *** represents significance at 1%.

### Figure 1. Fan Ratings and Total Points Scored

Source: PWW (2011)
Case study: 2012 NFL Fan Ratings

Fan ratings from NFL.com are available from all 256 NFL games played during the regular season in 2012. Data are also available on the following potential predictor variables:

- **MarginOfVictory**, the difference between the scores of the two teams
- **CombinedScore**, the combined score of the two teams
- **SumOfTeamRankings**, the sum of the two teams NFL.com Power Rankings rankings prior to the start of each game
- **OverTime**, a dummy variable which is 1 if the game goes into overtime
- **DivisionGame**, a dummy variable which is 1 if the game involves 2 teams from the same division
- **LateSundayAfternoon**, a dummy variable which is 1 if the game starts at 4pm or 4:25pm on Sunday

Apart from SumOfTeamRankings, the available predictor variables match those reported in Table 2 of PWW (2011). The predictor SumOfTeamRankings is to be used in place of “Sum of win percentage”, since the later does not take account of the difficulty of schedule.
2012 NFL Fan Ratings

Fitted model is as follows:
FanRating = 78.76 + 8.76 (if Thursday = 0) + 7.78 (if Overtime = 1) + 0.881CombinedScore (if CombinedScore<57) - 0.477SumOfTeamRankings (if SumOfTeamRankings>17) - 0.428MarginOfVictory (if MarginOfVictory < 16)
2012 NFL Fan Ratings

Fitted model is as follows:
FanRating = 78.76 + 8.76 (if Thursday = 0) + 7.78 (if Overtime = 1) + 0.881CombinedScore (if CombinedScore<57) - 0.477SumOfTeamRankings (if SumOfTeamRankings>17) - 0.428MarginOfVictory (if MarginOfVictory < 16)

Comparing this model with model II in Table 2 we see that
• The coefficients of CombinedScore and MarginOfVictory are the same sign in both models but otherwise quite different
• The coefficients of Overtime are similar
• The biggest difference is that all the effects in model II are linear.
NYC Taxi Trip Data

> 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

Air fare data

N=11,068,586

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

$$Q_M(\theta) = \sum_{i=1}^{n} \rho \left( \frac{y_i - \theta}{\sigma} \right)$$

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.
NYC Taxi Trip Data

The median\(\text{fare}_\text{amount}\) is a linear function of \(\text{rounded}_\text{trip}_\text{distance}\). In particular,

\[
\text{median}(\text{fare}_\text{amount}) = 2.50 + 2.50 \times \text{rounded}_\text{trip}_\text{distance} \quad (1)
\]

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.
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

$$Y = \beta_0 + \beta_1 (1500 - x)_- + \beta_2 (x - 1500)_+$$

(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

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".

Big data and precision

By D. R. Cox
Nuffield College, Oxford OX1 1NF, U.K.
david.cox@nuffield.ox.ac.uk
Taxi Trips - Dashboard

Taxi trips reported to the City of Chicago in its role as a regulatory agency. For the full data, see the bottom of this page or https://data.cityofchicago.org/di/wrvz-psew.

What's in this Dataset?

<table>
<thead>
<tr>
<th>Rows</th>
<th>Columns</th>
<th>Each row is a</th>
</tr>
</thead>
<tbody>
<tr>
<td>109M</td>
<td>23</td>
<td>Trip</td>
</tr>
</tbody>
</table>

Showing all Trips

Number of Trips by Trip Start Timestamp
When the trip started, rounded to the nearest 15 minutes.

Sum of Fares by Trip Start Timestamp
When the trip started, rounded to the nearest 15 minutes.
Monthly Chicago Taxi Fare Totals per Medallion over Time
Monthly Chicago Taxi Fare Totals over Time – 75th percentile

| Parameter | Estimate | Standard Error | t Value | Pr > |t| | Lag Variable | Shift |
|-----------|----------|----------------|---------|-------|---------------|-------------|-------|
| MU        | $8332.60 | 317.94         | 26.21   | <.0001| 0Quantile75_Monthly_Fare_Total_ | 0           |
| AR1,1     | 0.56     | 0.11           | 5.07    | <.0001| 1Quantile75_Monthly_Fare_Total_ | 0           |
| AR2,1     | 0.85     | 0.06           | 14.61   | <.0001| 12Quantile75_Monthly_Fare_Total_ | 0           |
| NUM1      | -$53.37  | 7.04           | -7.59   | <.0001| 0Months_since_2013              | 0           |
## Monthly Fare Totals over Time – 25th, 50th & 75th percentiles

| Parameter       | Estimate | Standard Error | t Value | Approx Pr > |t| | LagVariable                  |
|-----------------|----------|----------------|---------|-------------|-------------|-------------------------------|
| MU              | $8332.60 | 317.94         | 26.21   | <.0001      | 0Quantile75_Monthly_Fare_Total_|
| AR1,1           | 0.56     | 0.11           | 5.07    | <.0001      | 1Quantile75_Monthly_Fare_Total_|
| AR2,1           | 0.85     | 0.06           | 14.61   | <.0001      | 12Quantile75_Monthly_Fare_Total_|
| NUM1            | -$53.37  | 7.04           | -7.59   | <.0001      | 0Months_since_2013            |

| Parameter       | Estimate | Standard Error | t Value | Approx Pr > |t| | LagVariable                  |
|-----------------|----------|----------------|---------|-------------|-------------|-------------------------------|
| MU              | $5957.40 | 282.03         | 21.12   | <.0001      | 0Median_Monthly_Fare_Total_   |
| AR1,1           | 0.55     | 0.12           | 4.62    | <.0001      | 1Median_Monthly_Fare_Total_   |
| AR2,1           | 0.74     | 0.10           | 7.71    | <.0001      | 12Median_Monthly_Fare_Total_  |
| NUM1            | -$33.73  | 7.74           | -4.36   | <.0001      | 0Months_since_2013            |

| Parameter       | Estimate | Standard Error | t Value | Approx Pr > |t| | LagVariable                  |
|-----------------|----------|----------------|---------|-------------|-------------|-------------------------------|
| MU              | $3546.00 | 259.40         | 13.67   | <.0001      | 0Quantile25_Monthly_Fare_Total_|
| AR1,1           | 0.55     | 0.13           | 4.34    | <.0001      | 1Quantile25_Monthly_Fare_Total_|
| AR2,1           | 0.36     | 0.15           | 2.33    | 0.0196      | 12Quantile25_Monthly_Fare_Total_|
| NUM1            | -$31.96  | 8.87           | -3.6    | 0.0003      | 0Months_since_2013            |
Modeling exchange rate as a function of oil price

In December 2015, the FRED (Federal Reserve Bank of St. Louis) Blog posted a story entitled “The Canadian dollar and the price of oil” which says in part the following:

Canada’s oil sector amounts to about 10% of its GDP and 25% of its exports, almost all of which go to the U.S. It’s not too surprising, then, that the U.S./Canada exchange rate mirrors the price of oil. Of course, trade between the countries is much more than oil, but many of Canada’s other commodity exports have a price that is well correlated with the price of oil. And the financial linkages between the countries are also disproportionately tied to the mining and extractive industries.

Source: https://fredblog.stlouisfed.org/2015/12/the-canadian-dollar-and-the-price-of-oil/

We consider the monthly oil price and the US Canada exchange rate data obtained from the FRED from 1/1/2006 until 10/1/2016. In particular, we shall focus on the following two time series:

$X_t$, Oil price – Monthly crude oil price per barrel (West Texas Intermediate, Cushing, Oklahoma in $US$)


In this question, we wish to build a transfer function model in which $Y_t$, Exchange Rate CA$ in US$ is modeled as a function of $X_t$, Oil Price.
Modeling exchange rate as a function of oil price
Modeling exchange rate as a function of oil price
Modeling exchange rate as a function of oil price
Modeling exchange rate as a function of oil price
Modeling exchange rate as a function of oil price

Ignoring the MA error term, transfer function model 2 predicts that oil would have to increase by slightly more than $56 in price in a single month for Exchange Rate CA$ in US$ to increase by 0.1 or higher.
Examples of Work Based Capstone Projects - Class of 2017

• How Decision Trees Can Help Identify Fraud Patterns in Social Security SSI Disability Claims
• Predictive Sequential Association Rule Mining for Transactional Clickstream Data
• Predicting Bandwidth Utilization on Telecom Cell Towers
• Predicting Sales of Women’s Athletic Apparel
• Which Aspects of an Online Article Drive its Popularity
• Predicting Vehicle Crashes on Highways Ahead of Time
• Modeling the Relationship between Earned Media Activity and Service Engagement – Citi Bike NYC
• Predicting Market Rates for Drilling Rigs
• Times Series Analysis of US Rig Counts to Produce a Continuous Weekly Rig Count Prediction with a 12 Week Lead Time
Times Series Analysis of US Rig Counts to Produce a Continuous Weekly Rig Count Prediction with a 12 Week Lead Time

• In this study we shall focus on data for U.S. Rig Counts taken Baker Hughes from the years 2008 to 2017. In particular, we shall consider \( n = 440 \) weekly measurements of total rig count.

• Analysis was performed by Real Rig a start-up company that grew out of the Texas A&M University Analytics Program.

• Objective: To accurately predict the next quarter rig count on a weekly rolling basis.

• Transfer Function Model:
  \[
  \text{Rig Count}_{(t)} = \alpha \text{RigCount}_{(t-n)} + \beta \text{InputX}_{(t-q)} + \omega \text{InputY}_{(t-w)} \ldots + \xi
  \]

Email: info@realrig.com
Real Rig U.S. Rig Forecast

June 30, 2017

<table>
<thead>
<tr>
<th></th>
<th>Real Rig</th>
<th>Implied Upside*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gas</td>
<td>177</td>
<td>9%</td>
</tr>
<tr>
<td>Oil</td>
<td>773</td>
<td>13%</td>
</tr>
<tr>
<td>Total</td>
<td>950</td>
<td>12%</td>
</tr>
</tbody>
</table>

*Upside from 4/13/2017  Actual Source: BHI US Rig Count

Email: info@realrig.com
Comparison to Industry

Quarterly Average of Total Rigs

- Actual
- Real Rig
- Leading Industry Forecast

Outperforming leading industry forecast

<table>
<thead>
<tr>
<th>Quarter</th>
<th>Actual</th>
<th>Real Rig</th>
<th>Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>4Q'16</td>
<td>589</td>
<td>551</td>
<td>529</td>
</tr>
<tr>
<td>1Q'17</td>
<td>742</td>
<td>732</td>
<td>646</td>
</tr>
<tr>
<td>2Q'17</td>
<td>843*</td>
<td>898</td>
<td>885</td>
</tr>
</tbody>
</table>

Percent Error

<table>
<thead>
<tr>
<th>Quarter</th>
<th>Real Rig</th>
<th>Industry</th>
<th>Decreased Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>4Q'16</td>
<td>-6%</td>
<td>-10%</td>
<td>37%</td>
</tr>
<tr>
<td>1Q'17</td>
<td>-1%</td>
<td>-13%</td>
<td>90%</td>
</tr>
</tbody>
</table>

*Average as of 4/13/2017  Actual Source : BHI US Rig Count

Email : info@realrig.com
Questions