ANALYTICS
TEXAS A&M UNIVERSITY

Flexible Statistical Modeling Methods for Big Data

April 21, 2017

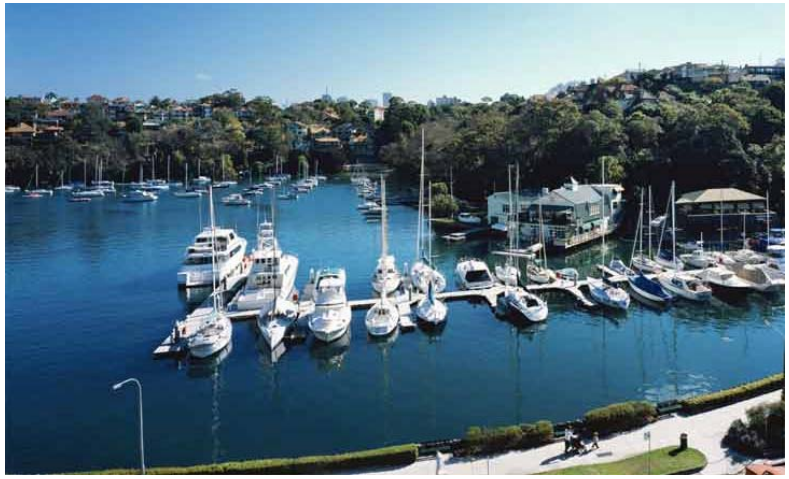
Professor Simon Sheather **email: sheather@stat.tamu.edu**

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.
- Henry (1797-1865) > Reuben (1827) > James (1890-1959) > Kevin (1925-2014)> Simon

The 4 Sheather Brothers
Which one am I?

The 4 Sheather Brothers
Which one am I?

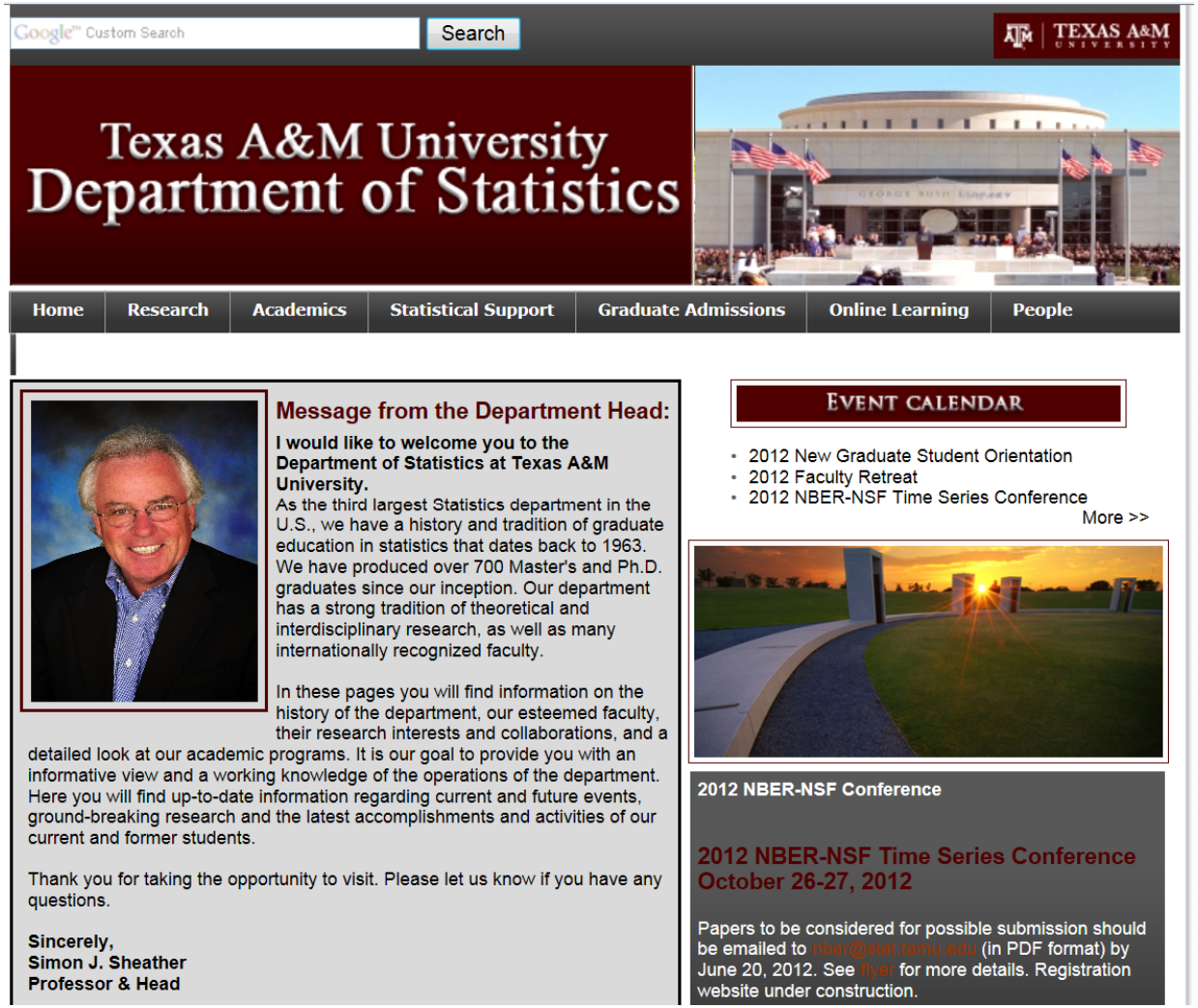
Andrew, Contractor

Simon, Professor

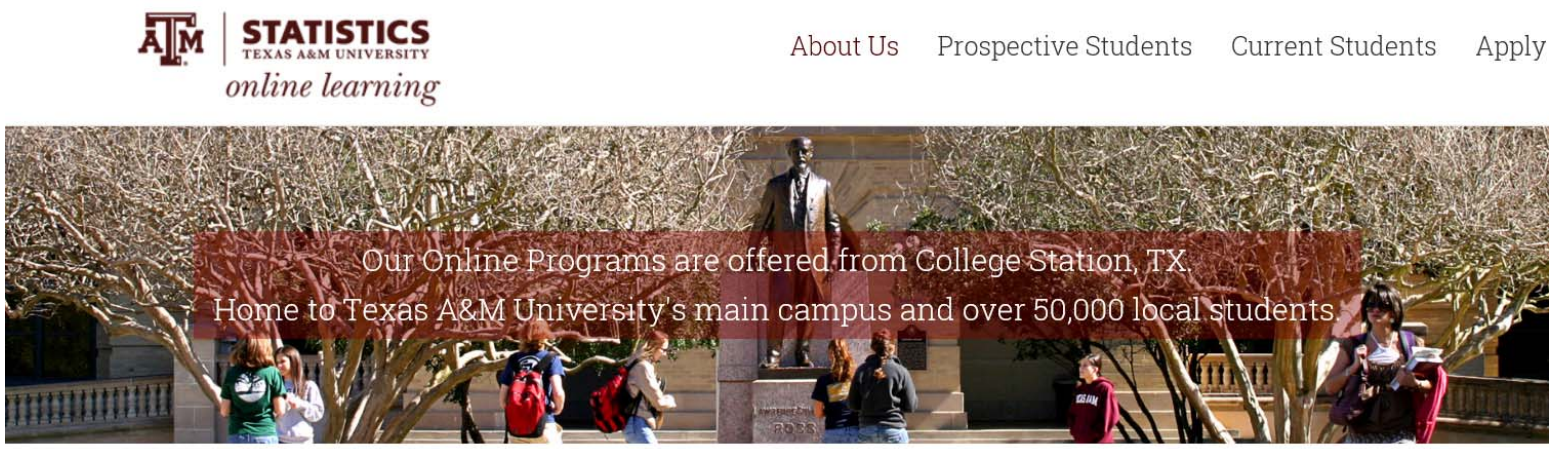
Martin, Banker

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



Texas A&M Statistical Services LP was formed in 2012

TEXAS A&M
STATISTICAL SERVICES, LP

OUR SERVICES

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

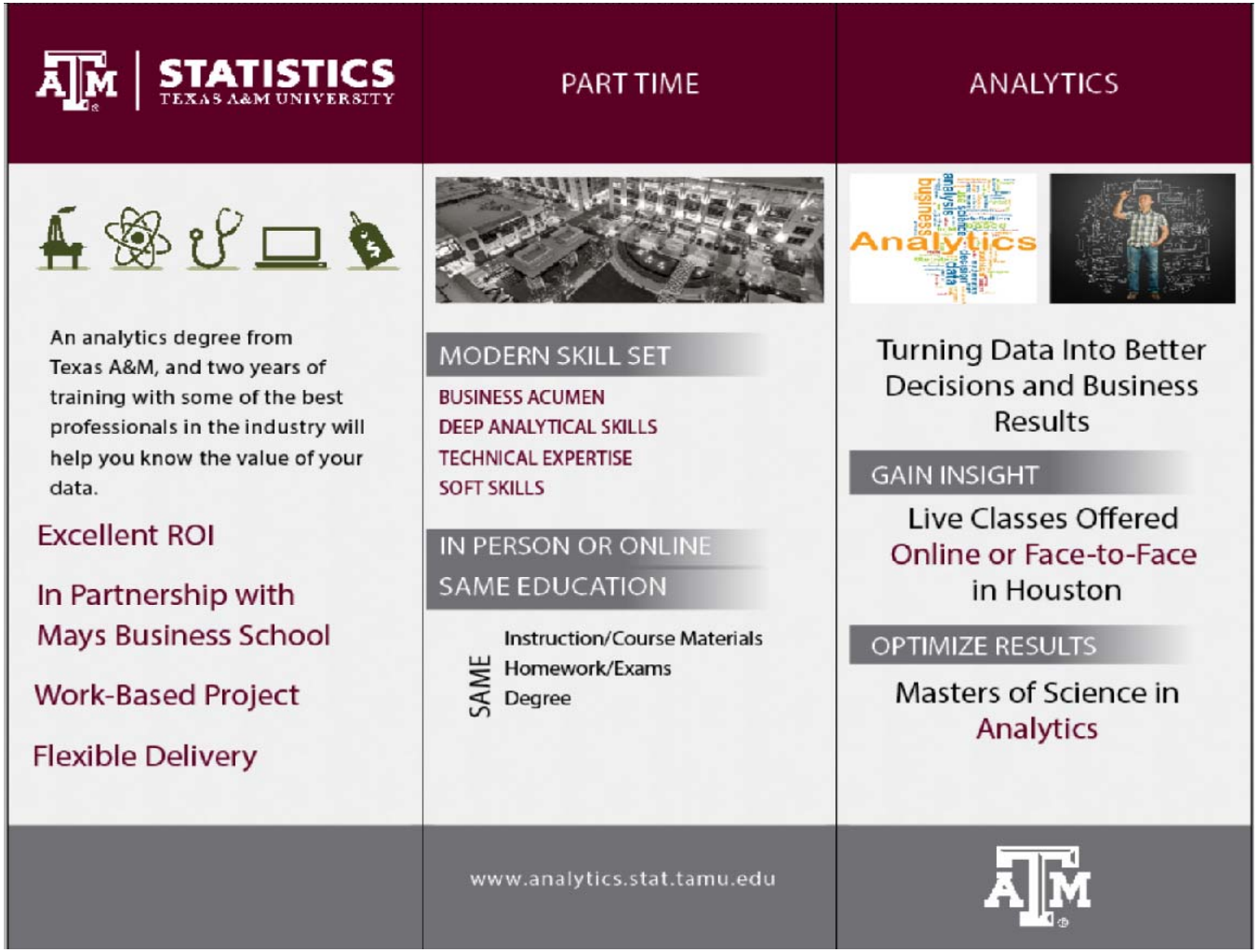
We are dedicated to helping organizations harness data and apply analytics to product design, service improvement, marketing and decison-making.

Our services include:

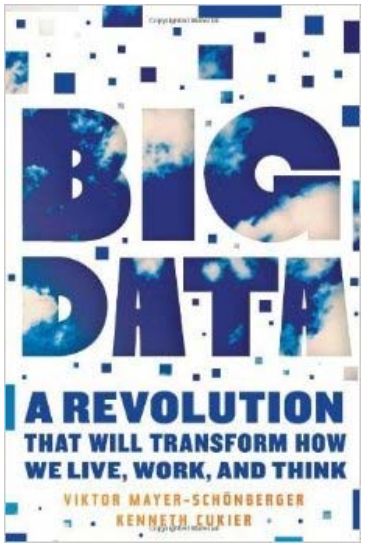
- **Business Assessments for Analytics**
- **Public Webinars & Workshops**
- **Customized Corporate Training**
- **Analytics Support for Mission Critical Projects**
- **Customer Satisfaction Analysis & Tracking**

http://www.tamstatservices.com/

In Fall 2013, MS (Analytics) program in partnership with the Mays Business School

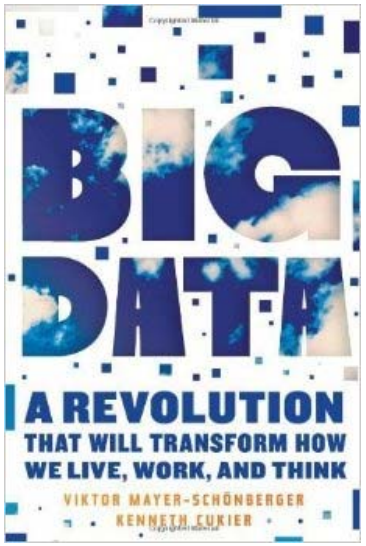


Definitions of “big data”



“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* = *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 *Sophisticatedly 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 Sophisticatedly 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".**

Simplicity, Inference and Modelling

Keeping it Sophisticatedly Simple

edited by

Arnold Zellner, Hugo A. Keuzenkamp and Michael McAleer

CAMBRIDGE UNIVERSITY PRESS

Predictive Analytics

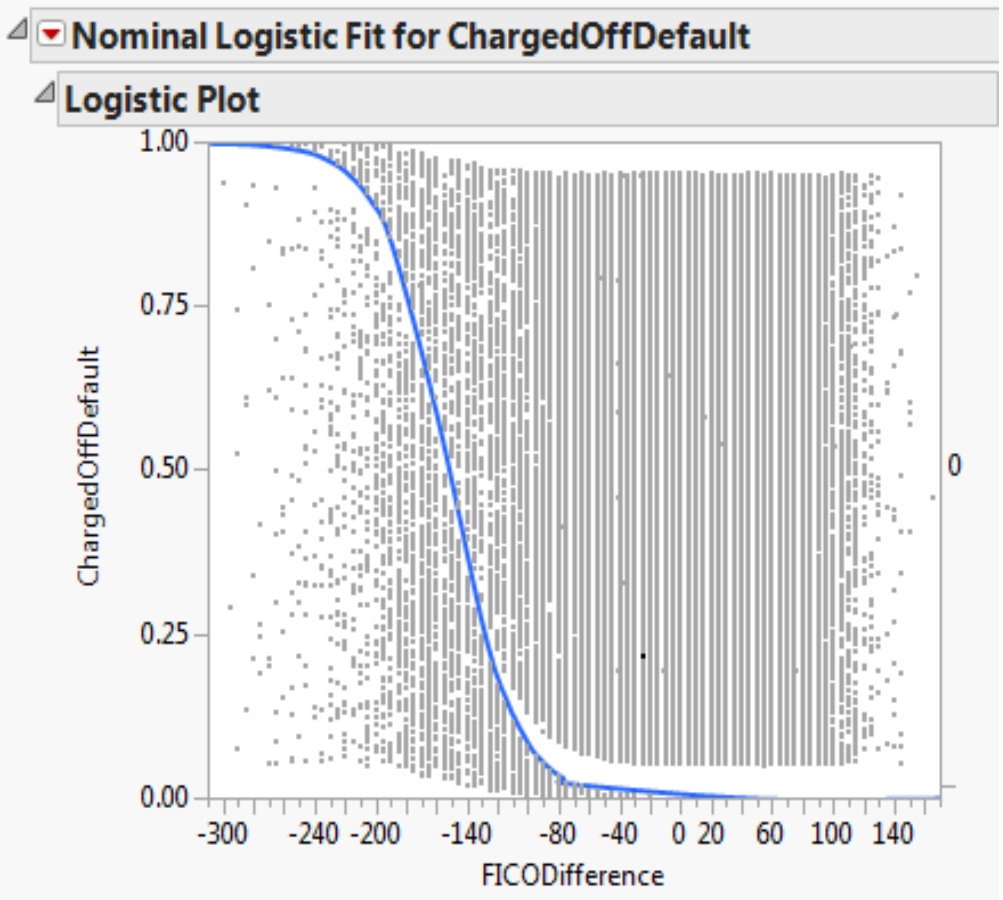
"Predictive analytics encompasses a variety of techniques from statistics, modeling, machine learning, and data mining that analyze current and historical facts **to make predictions** about future, or otherwise unknown, events

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

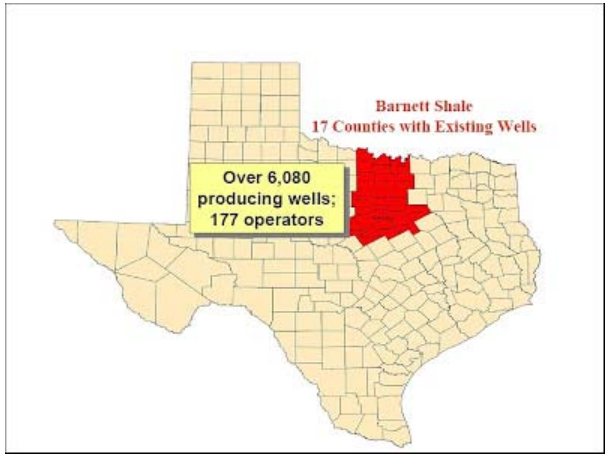
Subject Matter Expertise is Important in Model Development



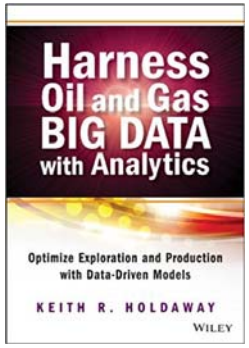
ANALYTICS
TEXAS A&M UNIVERSITY

Modeling 12 month non-zero gas production

- Interest centers on developing a model for 12 month non-zero gas production in unconventional reservoirs in the Barnett Shale using data from around 5000 wells, taken from Holdaway (2014, *Harness Oil and Gas Big Data with Analytics*, Wiley).
- 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/



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

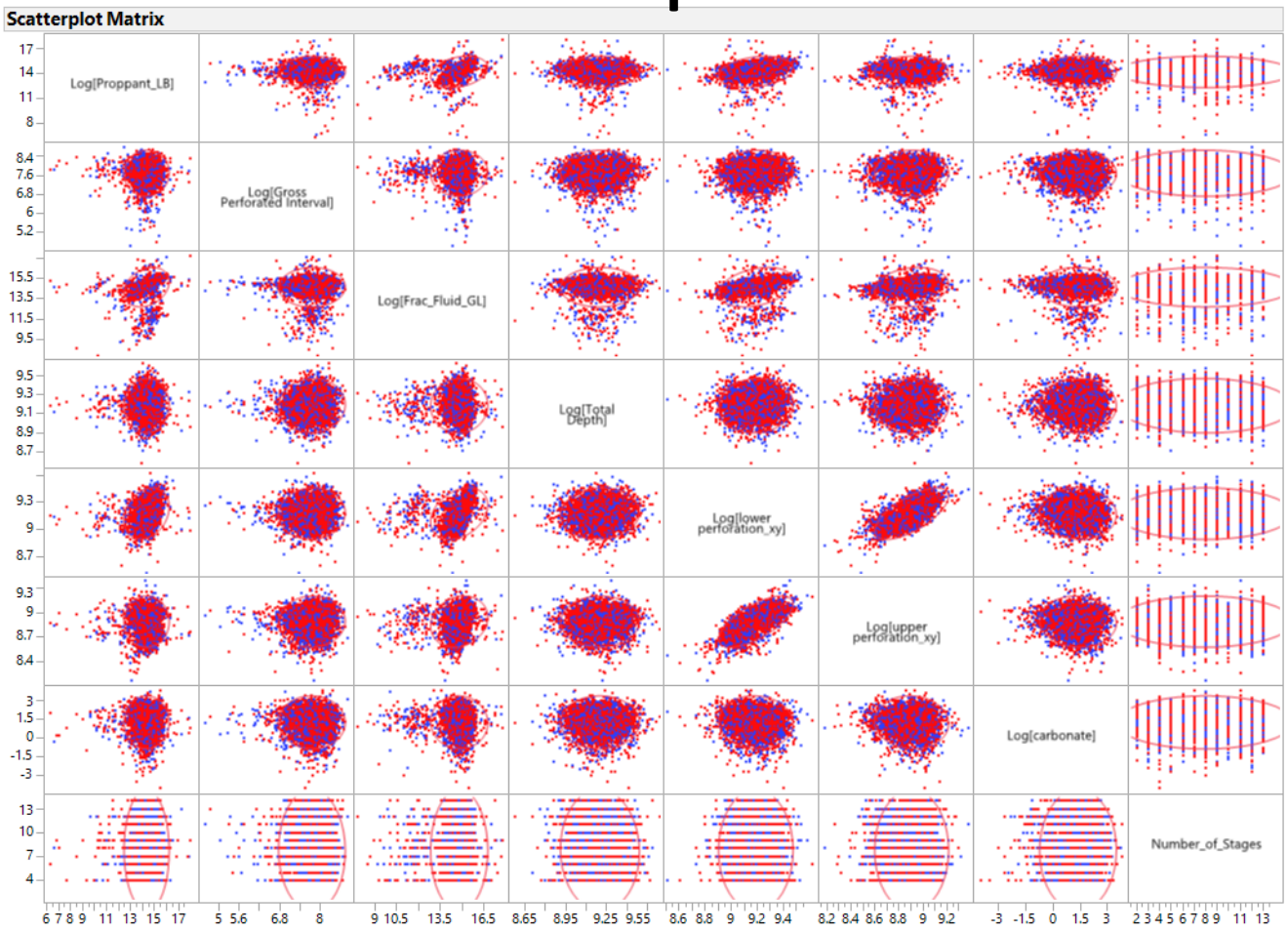
Modeling 12 month non-zero gas production

Correlations

	Log[Proppant_LB]	Log[Gross Perforated Interval]	Log[Frac_Fluid_GL]	Log[Total Depth]	Log[lower perforation_xy]	Log[upper perforation_xy]	Log[carbonate]	Number_of_Stages
Log[Proppant_LB]	1.0000	0.0153	0.2746	0.0242	0.2659	0.0018	0.0076	-0.0140
Log[Gross Perforated Interval]	0.0153	1.0000	0.0227	0.0325	0.0345	0.0211	-0.0131	-0.0133
Log[Frac_Fluid_GL]	0.2746	0.0227	1.0000	0.0235	0.3234	0.1713	-0.0263	-0.0027
Log[Total Depth]	0.0242	0.0325	0.0235	1.0000	0.0588	0.0168	-0.0016	-0.0255
Log[lower perforation_xy]	0.2659	0.0345	0.3234	0.0588	1.0000	0.6829	-0.0501	0.0009
Log[upper perforation_xy]	0.0018	0.0211	0.1713	0.0168	0.6829	1.0000	-0.0532	0.0021
Log[carbonate]	0.0076	-0.0131	-0.0263	-0.0016	-0.0501	-0.0532	1.0000	0.0051
Number_of_Stages	-0.0140	-0.0133	-0.0027	-0.0255	0.0009	0.0021	0.0051	1.0000

- 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:

-
-

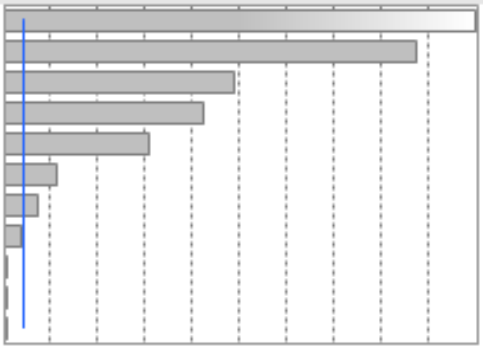
Modeling 12 month non-zero gas production – only first order terms

Response Log[nzd_12_cum]

Weight: Training

Effect Summary

Source	LogWorth	PValue
County	67.165	0.00000
Log[lower perforation_xy]	43.833	0.00000
X_Well	24.520	0.00000
Log[Proppant_LB]	21.327	0.00000
Log[upper perforation_xy]	15.432	0.00000
Log[Frac_Fluid_GL]	5.715	0.00000
Y_Well	3.725	0.00019
Number_of_Stages	1.859	0.01384
Log[carbonate]	0.401	0.39692
Log[Gross Perforated Interval]	0.301	0.49983
Log[Total Depth]	0.270	0.53686

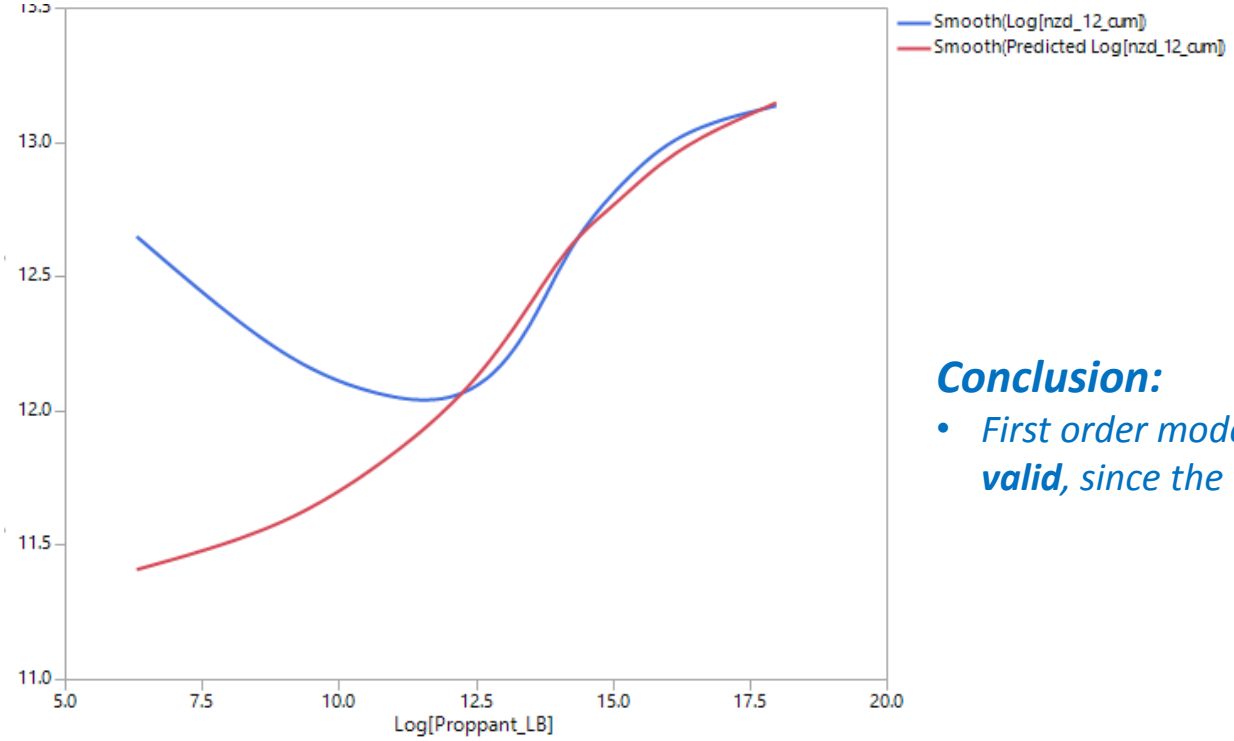


Summary of Fit

RSquare	0.428613
RSquare Adj	0.424509
Root Mean Square Error	0.520923
Mean of Response	12.59958
Observations (or Sum Wgts)	3226

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

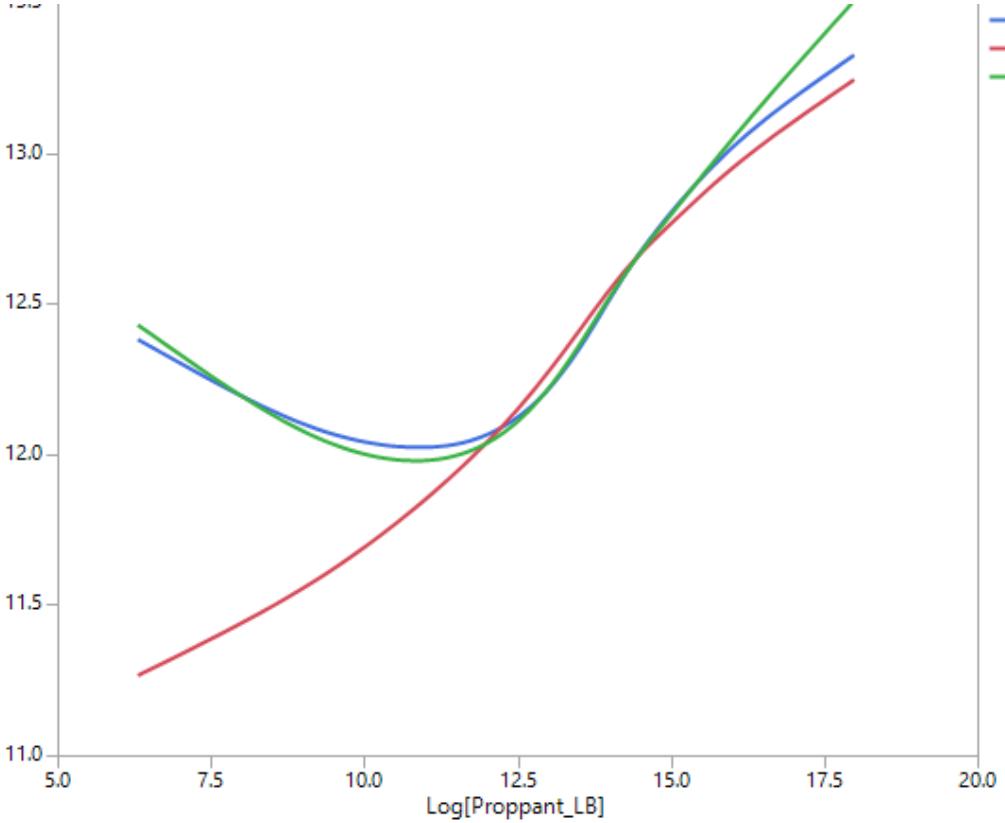
Response Log[nzd_12_cum]

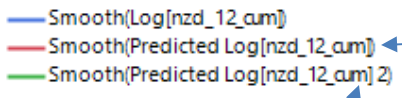
Weight: Training

Effect Summary

Source	LogWorth	PValue
County	55.396	0.00000
Log[Proppant_LB]	28.818	0.00000
Log[lower perforation_xy]	26.228	0.00000
Y_Well*Y_Well	26.145	0.00000
X_Well	15.769	0.00000
Log[Proppant_LB]*Log[Proppant_LB]	12.691	0.00000
Y_Well	9.595	0.00000
Log[Frac_Fluid_GL]	9.328	0.00000
Log[Frac_Fluid_GL]*Log[Frac_Fluid_GL]	7.430	0.00000
Log[upper perforation_xy]	6.502	0.00000
X_Well*X_Well	5.492	0.00000
Number_of_Stages	2.868	0.00135
Number_of_Stages*Number_of_Stages	2.630	0.00234
Log[Proppant_LB]*X_Well	2.029	0.00936
Log[Gross Perforated Interval]*Log[Total Depth]	1.998	0.01004
Log[Proppant_LB]*Log[lower perforation_xy]	1.952	0.01117
Log[lower perforation_xy]*Log[upper perforation_xy]	1.577	0.02651
Log[lower perforation_xy]*X_Well	1.548	0.02834
Log[Proppant_LB]*Y_Well	1.443	0.03602
Log[Frac_Fluid_GL]*Log[upper perforation_xy]	1.420	0.03801

Modeling non-zero 12 month gas production – Marginal model plot





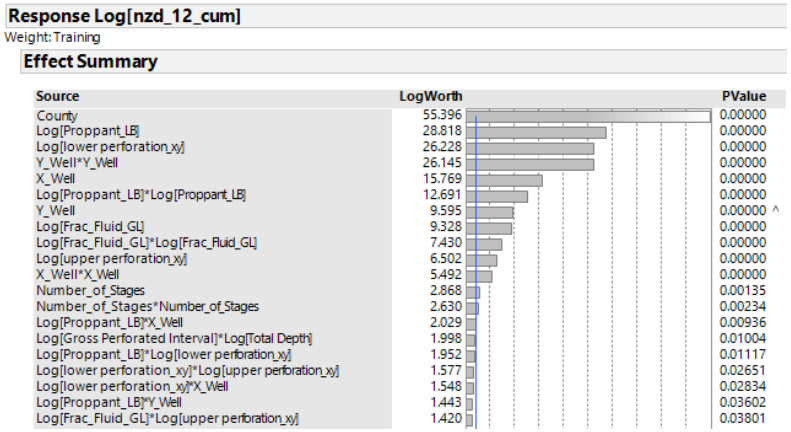
Only first order terms

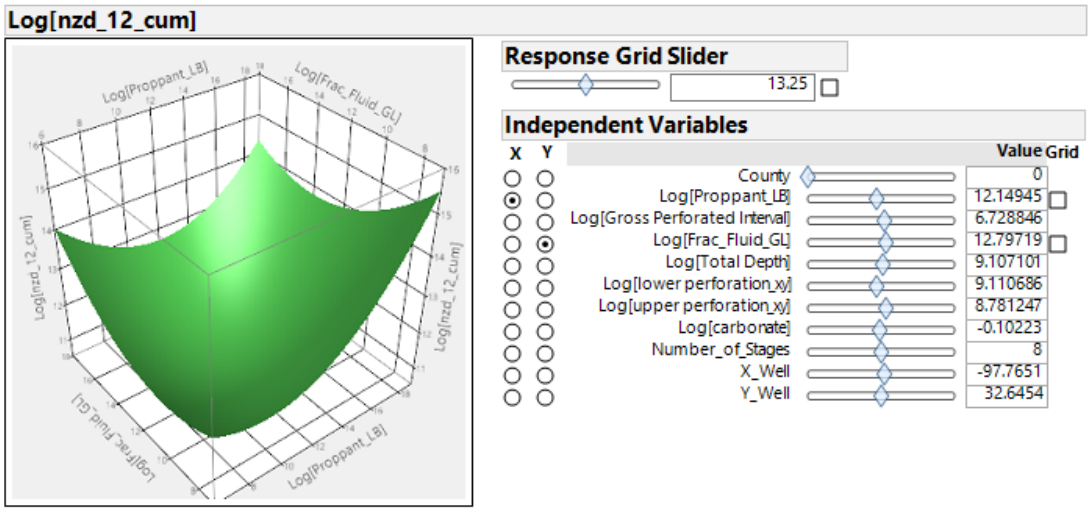
First & second order terms

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



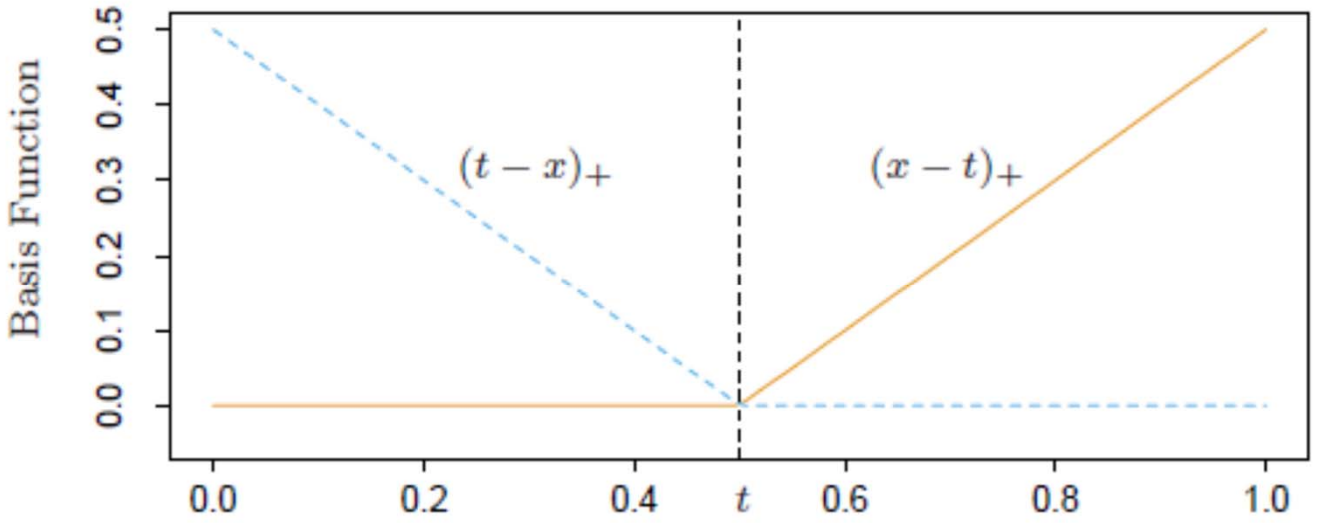


Recall that the initial focus was on both **proppant** and **fracture fluid volumes.** The second order model finds that Log[nzd_12_cum] is maximized for high values of **proppant** and low values of **fracture fluid volumes.**

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

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



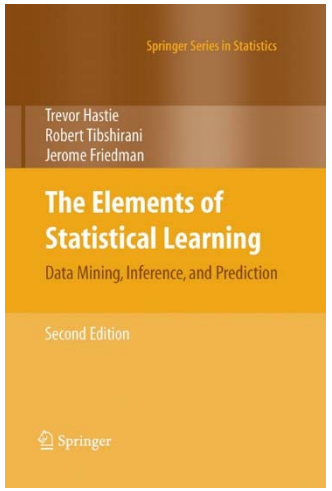
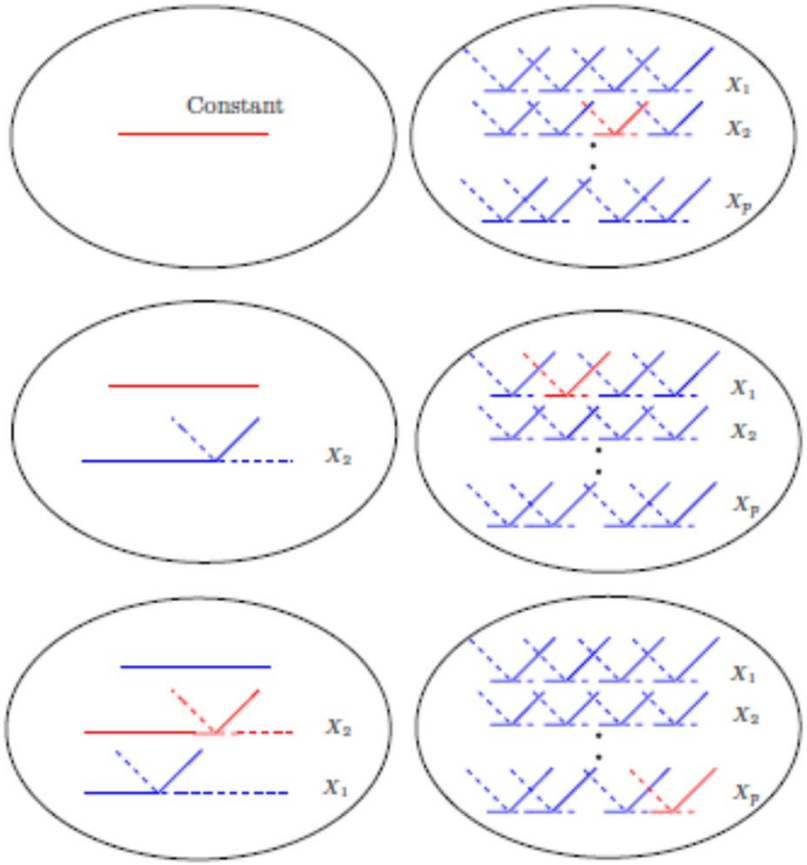


FIGURE 9.9. *The basis functions $(x-t)_+$ (solid orange) and $(t-x)_+$ (broken blue) used by MARS.*

MARS
(Multiple adaptive regression splines)



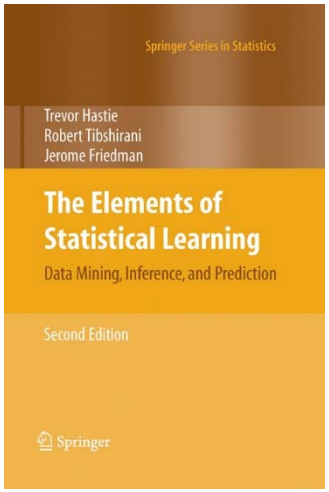


FIGURE 9.10. Schematic of the MARS forward model-building procedure. On the left are the basis functions currently in the model: initially, this is the constant function. On the right are all candidate basis functions to be considered in building the model. These are pairs of piecewise linear basis functions 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

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

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

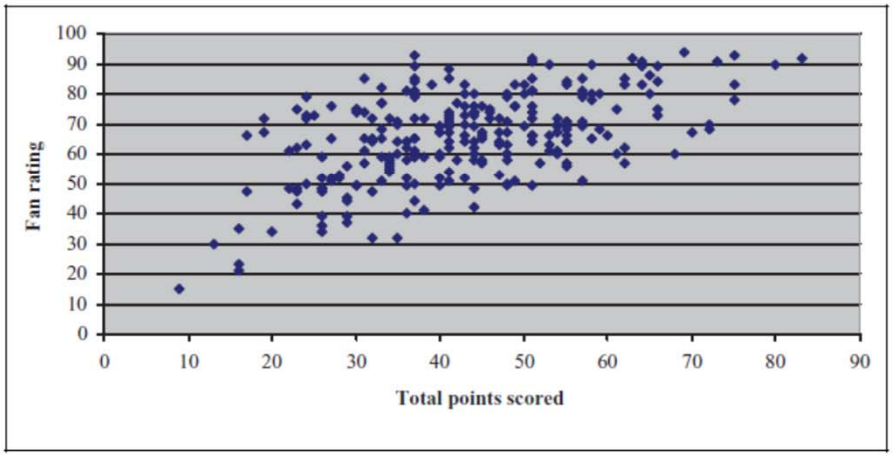


Figure 1. Fan Ratings and Total Points Scored

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)

Basis Information

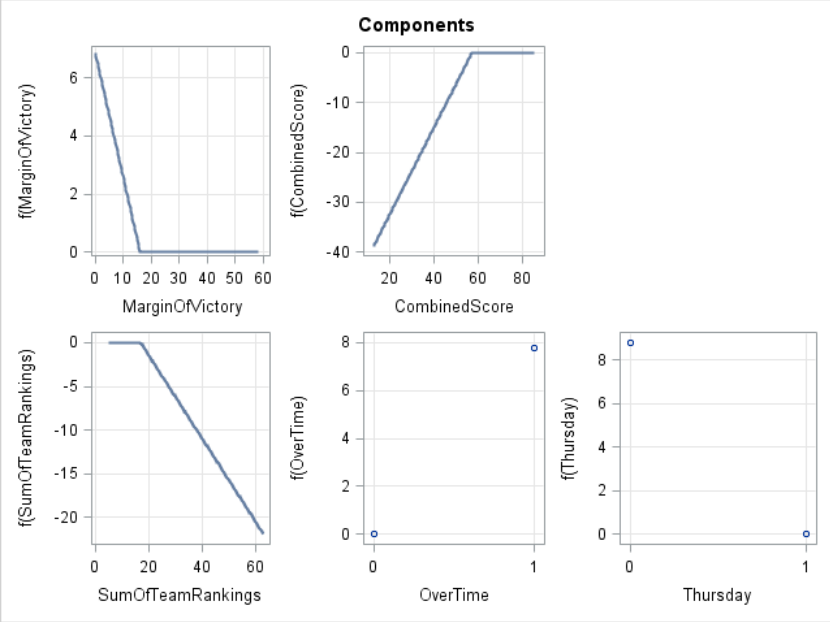
Name	Transformation
Basis0	1
Basis1	Basis0*MAX(CombinedScore - 57,0)
Basis2	Basis0*MAX(57 - CombinedScore,0)
Basis3	Basis0*MAX(SumOfTeamRankings - 17,0)
Basis4	Basis0*MAX(17 - SumOfTeamRankings,0)
Basis5	Basis0*MAX(MarginOfVictory - 16,0)
Basis6	Basis0*MAX(16 - MarginOfVictory,0)
Basis7	Basis0*(Thursday = 0)
Basis8	Basis0*NOT(Thursday = 0)
Basis9	Basis0*(OverTime = 1)
Basis10	Basis0*NOT(OverTime = 1)
Basis11	Basis0*MAX(CombinedScore - 30,0)
Basis12	Basis0*MAX(30 - CombinedScore,0)
Basis13	Basis0*MAX(CombinedScore - 33,0)
Basis14	Basis0*MAX(33 - CombinedScore,0)
Basis15	Basis0*MAX(SumOfTeamRankings - 53,0)
Basis16	Basis0*MAX(53 - SumOfTeamRankings,0)
Basis17	Basis0*MAX(SumOfTeamRankings - 16,0)
Basis18	Basis0*MAX(16 - SumOfTeamRankings,0)
Basis19	Basis0*MAX(CombinedScore - 69,0)
Basis20	Basis0*MAX(69 - CombinedScore,0)

Regression Spline Model after Backward Selection

Name	Coefficient	Parent	Variable	Knot	Levels
Basis0	78.7561		Intercept		
Basis2	-0.8810	Basis0	CombinedScore	57.0000	
Basis3	-0.4769	Basis0	SumOfTeamRankings	17.0000	
Basis6	0.4284	Basis0	MarginOfVictory	16.0000	
Basis7	8.7826	Basis0	Thursday		0
Basis9	7.7828	Basis0	OverTime		1

Variable Importance

Variable	Number of Bases	Importance
CombinedScore	1	100.00
SumOfTeamRankings	1	33.51
Thursday	1	5.81
MarginOfVictory	1	3.83
OverTime	1	3.30



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.

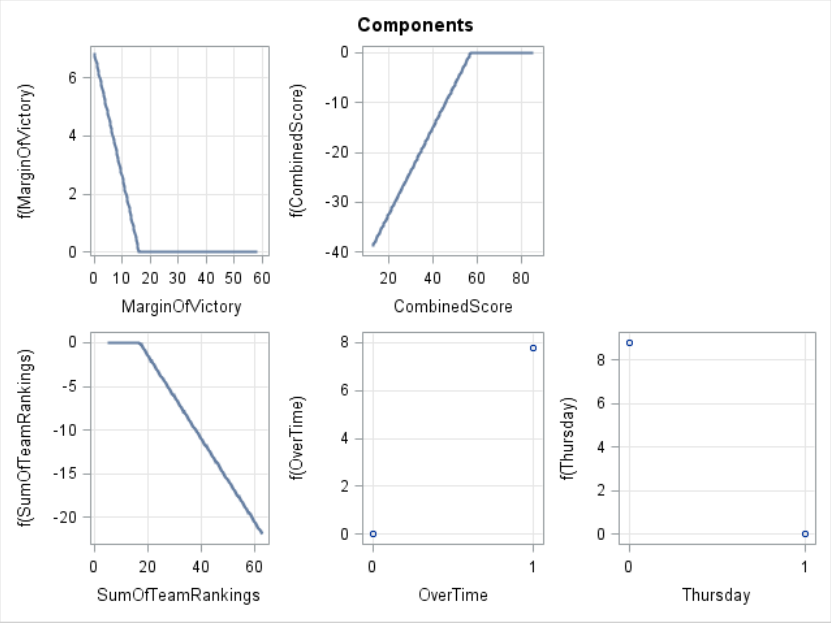


Table 2. Regression Results—Determinants of Fan Ratings of NFL Games

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

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

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

Trip Sheet Data (CSV Format)

2016

January	Yellow	Green	FHV
February	Yellow	Green	FHV
March	Yellow	Green	FHV
April	Yellow	Green	FHV
May	Yellow	Green	FHV
June	Yellow	Green	FHV
July	Yellow	Green	FHV
August	Yellow	Green	FHV
September	Yellow	Green	FHV
October	Yellow	Green	FHV
November	Yellow	Green	FHV
December	Yellow	Green	FHV

2015

2014

2013

2012

2011

2010

2009

Source: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

Air fare data

N=11,068,586

United States Department of Transportation

Office of the Assistant Secretary for Research and Technology

Bureau of Transportation Statistics

Data Profile: Airline Origin and Destination Survey (DB1B)

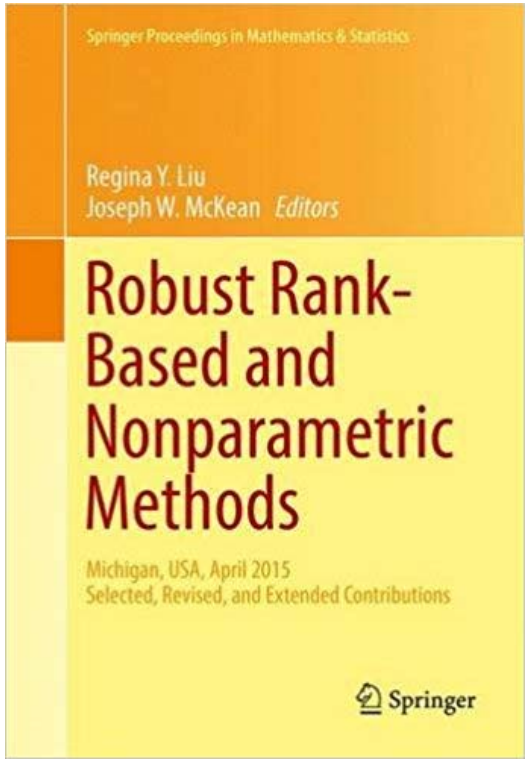
Printable Version Print | Data Tables

Airline Origin and Destination Survey (DB1B)

Overview

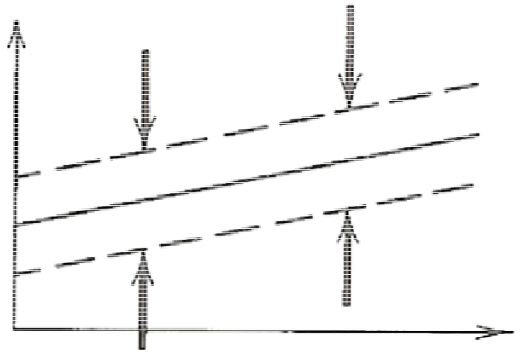
The Airline Origin and Destination Survey (DB1B) is a 10% sample of airline tickets from reporting carriers collected by the Office of Airline Information of the Bureau of Transportation Statistics. Data includes origin, destination and other itinerary details of passengers transported. This database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

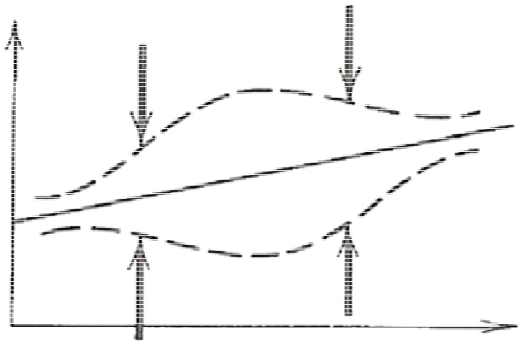
http://www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=125&DB_Name=Airline%20Origin%20and%20Destination%20Survey%20(DB1B)



Sheather, S.J. (2016) Applications of robust regression to "big" data problems, *Robust Rank-Based and Nonparametric Methods* Springer, New York, 101-120.

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 θ (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 θ (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 θ (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 θ and σ 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 θ (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

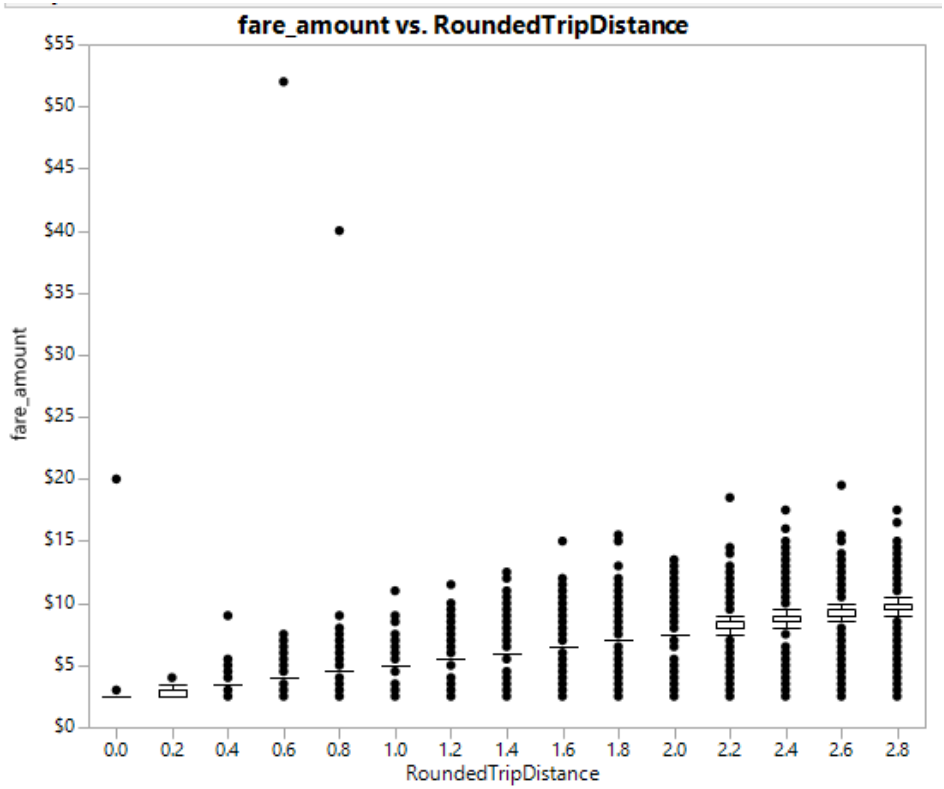
NYC Taxi Trip Data

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

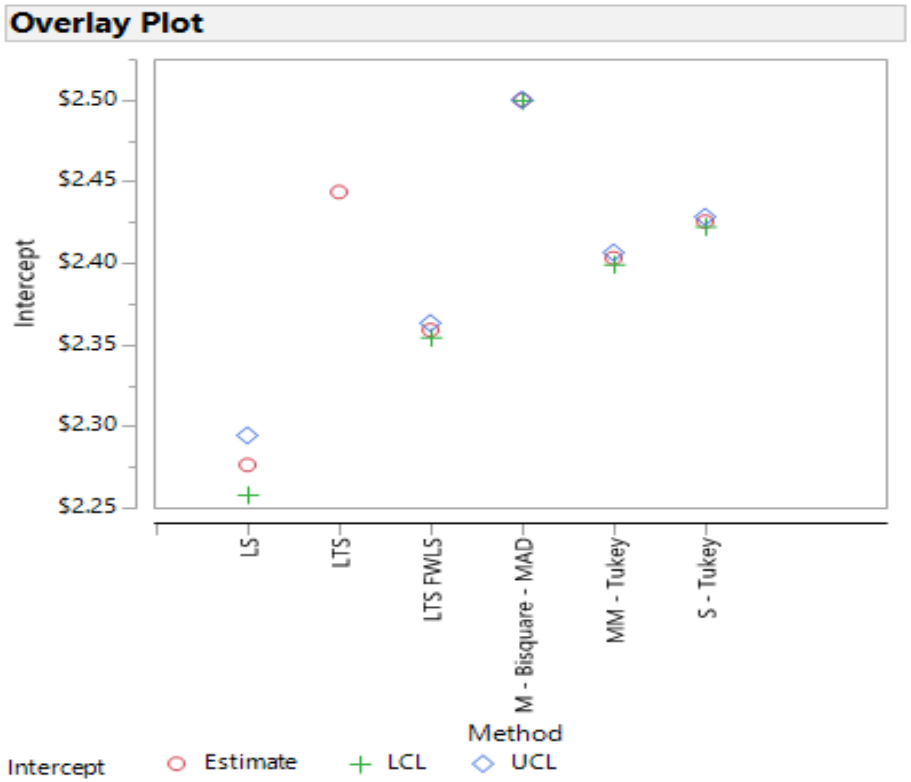


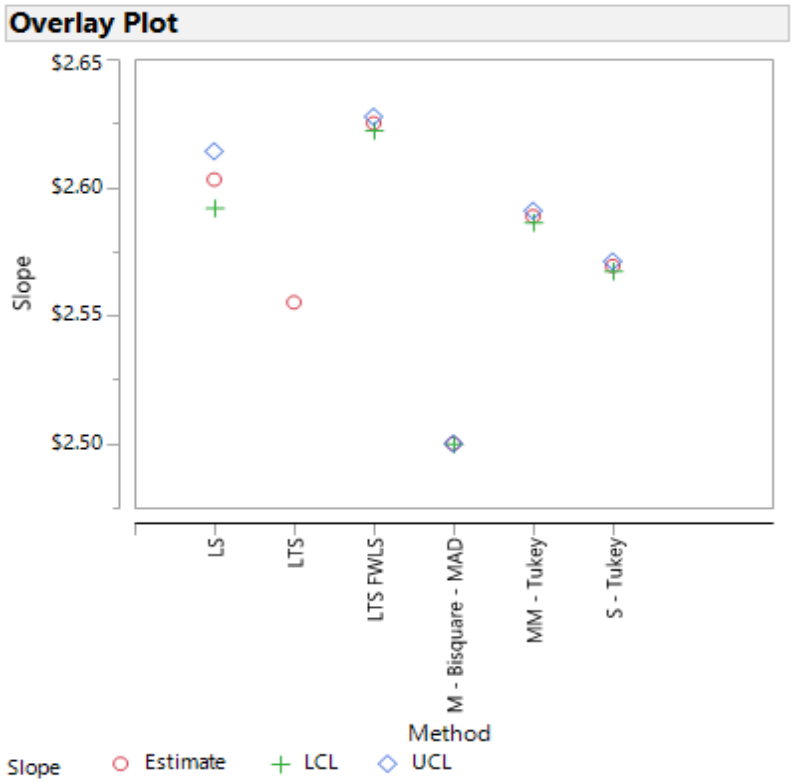
The median(fare_amount) is a linear function of rounded_trip_distance. In particular,

$$\text{median}(\text{fare\_amount}) = \$2.50 + \$2.50 * \text{rounded\_trip\_distance} \quad (1)$$

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



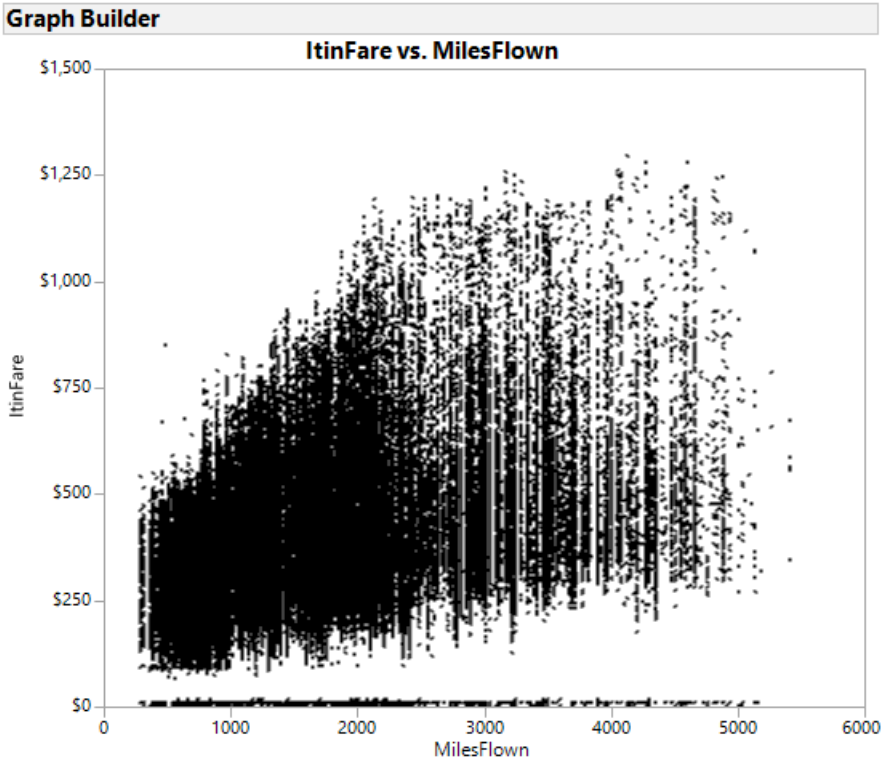


Conclusions:

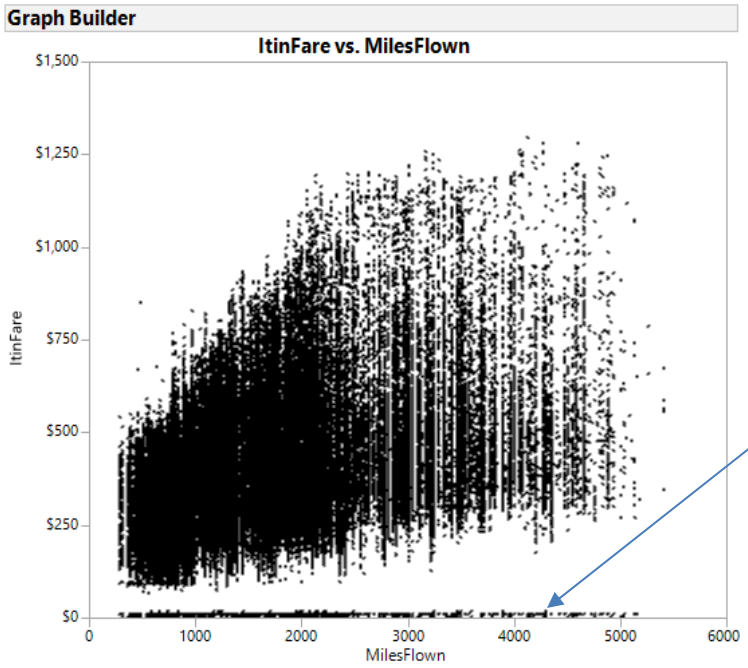
1. Only the M-estimates and the R-estimates are equal to the values of the intercept and the slope in (1), namely, $2.50.
2. The confidence intervals are very narrow implying high precision of the point estimates.

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.



Air fare data



Frequent flyer tickets

Denote ItinFare by Y and MilesFlown by x. We considered regression spline models of the form

$$Y = \beta_0 + \beta_1(1500 - x)_- + \beta_2(x - 1500)_+ \qquad (2)$$

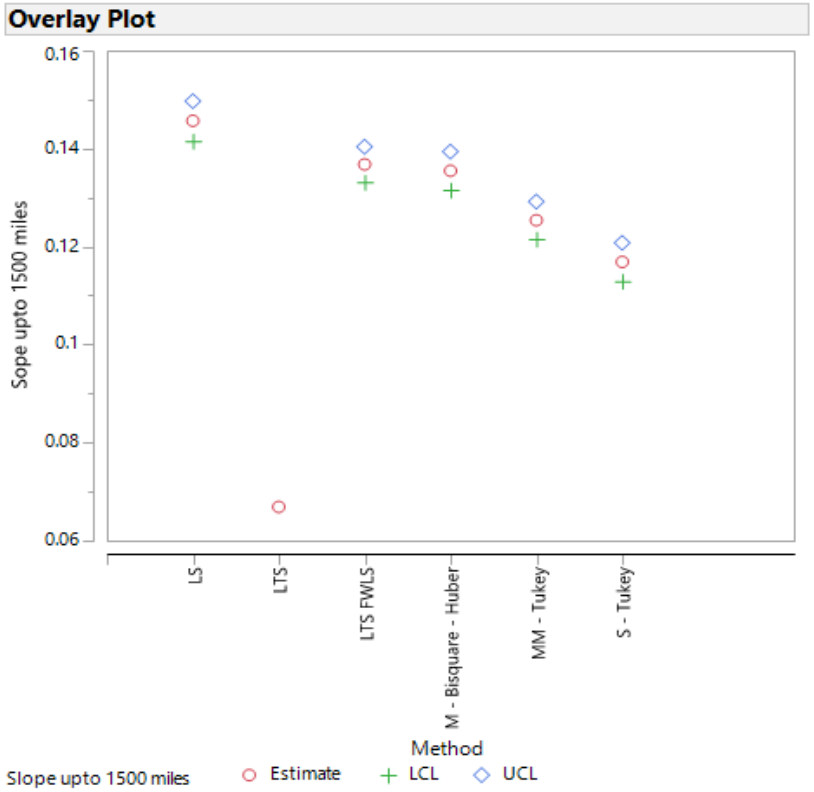
where

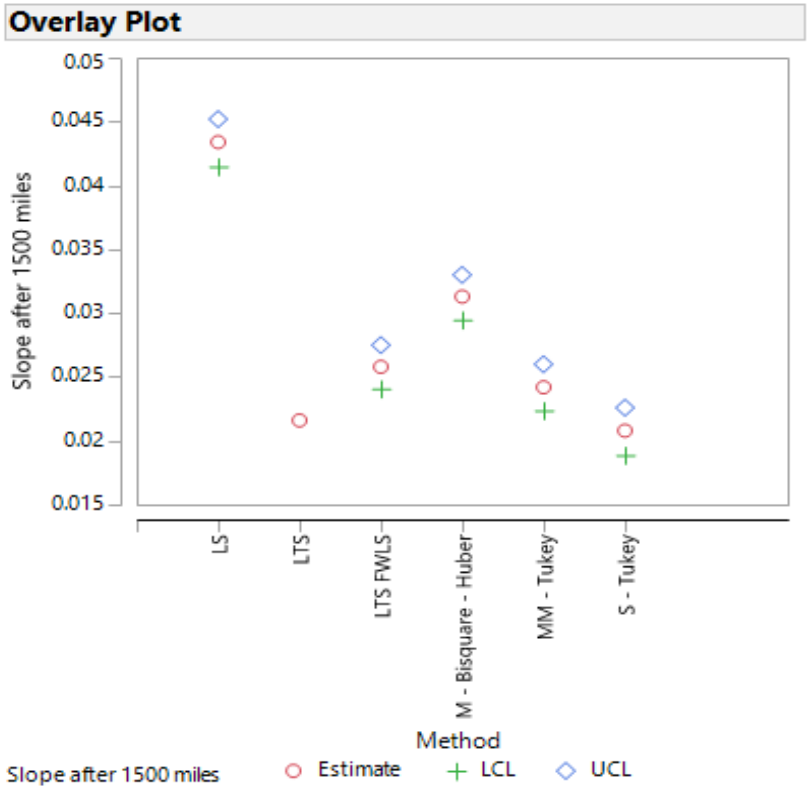
$$(1500 - x)_- = \begin{cases} x - 1500, & x < 1500 \\ 0, & x \geq 1500 \end{cases}$$

and

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

Air fare data





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".

Biometrika (2015), *pp.* 1–5
Printed in Great Britain

doi: 10.1093/biomet/asv033

Big data and precision

BY D. R. COX

Nuffield College, Oxford OX1 1NF, U.K.

david.cox@nuffield.ac.ox.uk

https://data.cityofchicago.org/Transportation/Taxi-Trips-Dashboard/spcw-brbq

Taxi Trips - Dashboard

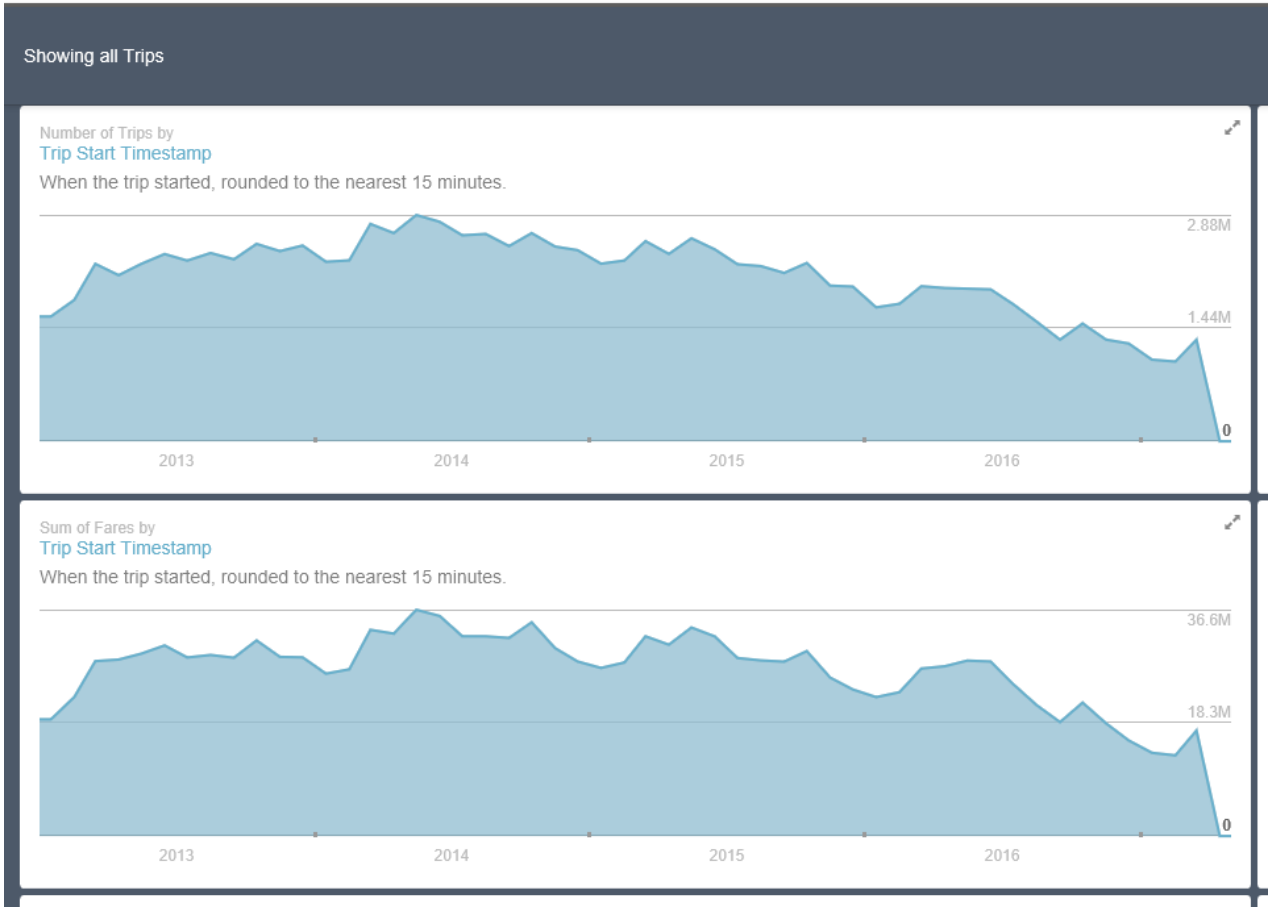
Taxi trips reported to the City of Chicago in its role as a regulatory agency. For the full data, see the bottom of this page or https://data.cityofchicago.org/d/wrvz-psew.
...

Show more

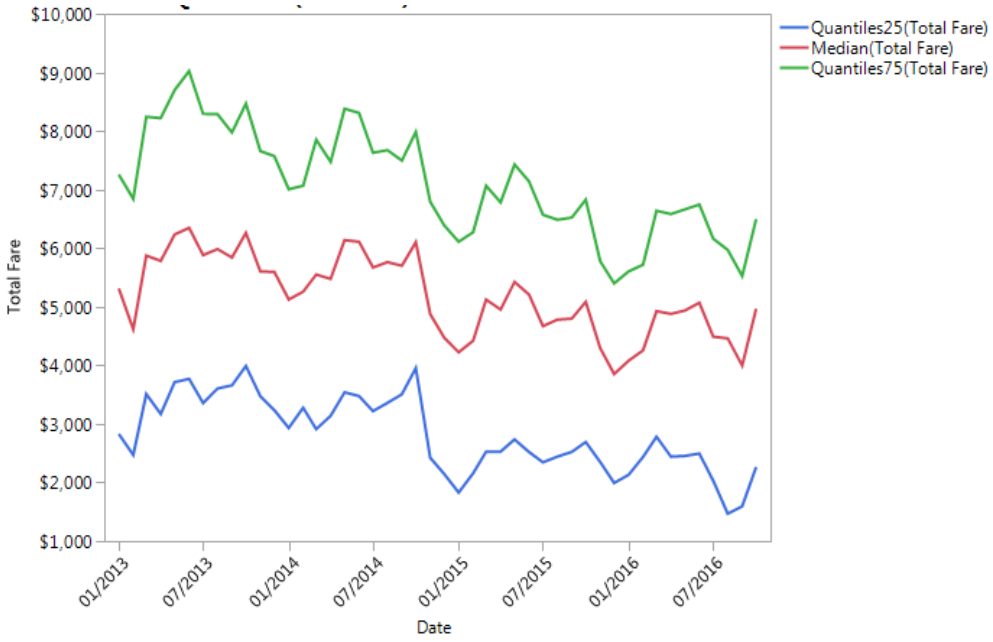
Export API

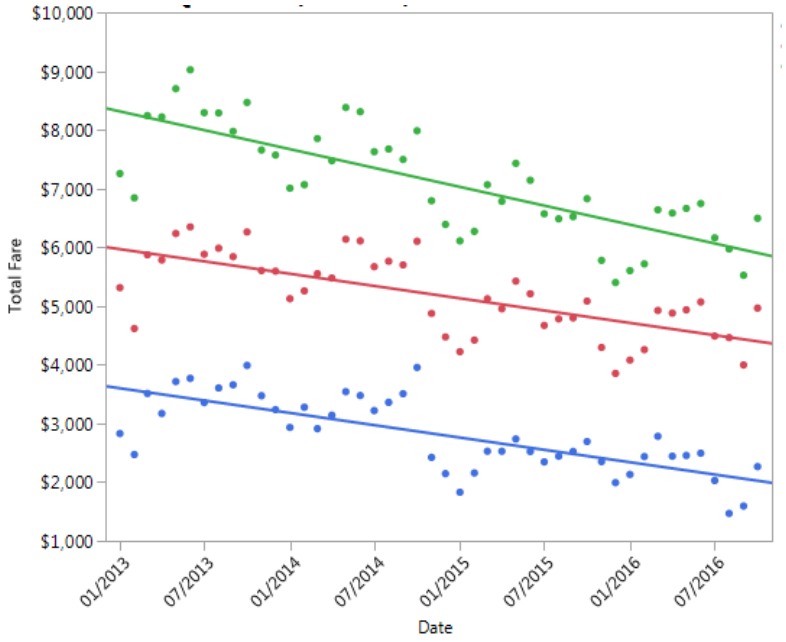
What's in this Dataset?

Rows	Columns	Each row is a
109M	23	Trip



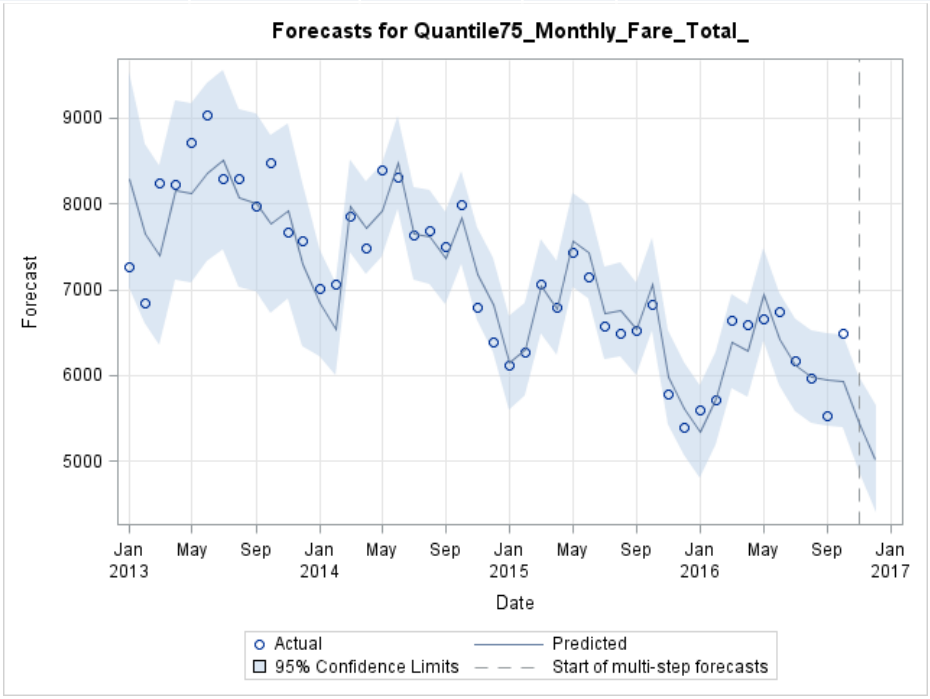
Monthly Chicago Taxi Fare Totals per Medallion over Time





Monthly Chicago Taxi Fare Totals over Time – 75th percentile

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



Monthly Fare Totals over Time – 25th, 50th & 75th percentiles

Maximum Likelihood Estimation						
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable
MU	**$8332.60**	317.94	26.21	<.0001	0	Quantile75_Monthly_Fare_Total_
AR1,1	0.56	0.11	5.07	<.0001	1	Quantile75_Monthly_Fare_Total_
AR2,1	0.85	0.06	14.61	<.0001	12	Quantile75_Monthly_Fare_Total_
NUM1	**-$53.37**	7.04	-7.59	<.0001	0	Months_since_2013
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable
MU	**$5957.40**	282.03	21.12	<.0001	0	Median_Monthly_Fare_Total_
AR1,1	0.55	0.12	4.62	<.0001	1	Median_Monthly_Fare_Total_
AR2,1	0.74	0.10	7.71	<.0001	12	Median_Monthly_Fare_Total_
NUM1	**-$33.73**	7.74	-4.36	<.0001	0	Months_since_2013
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable
MU	**$3546.00**	259.40	13.67	<.0001	0	Quantile25_Monthly_Fare_Total_
AR1,1	0.55	0.13	4.34	<.0001	1	Quantile25_Monthly_Fare_Total_
AR2,1	0.36	0.15	2.33	0.0196	12	Quantile25_Monthly_Fare_Total_
NUM1	**-$31.96**	8.87	-3.6	0.0003	0	Months_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)

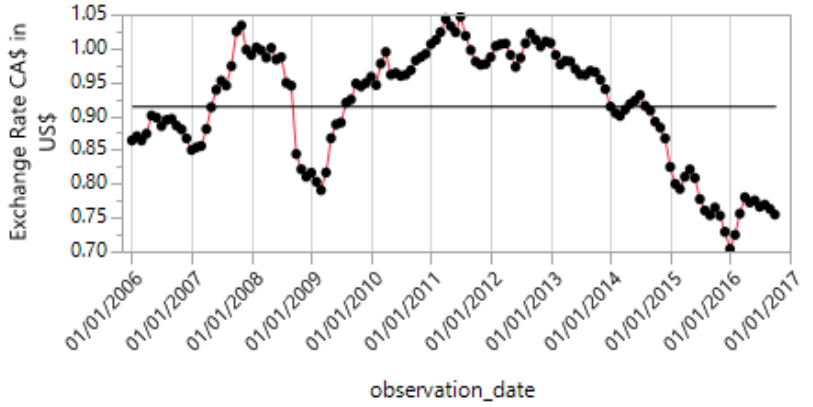
Y_t, Exchange Rate CA$ in US$ - Monthly exchange rate of the Canadian dollar in US dollars.

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

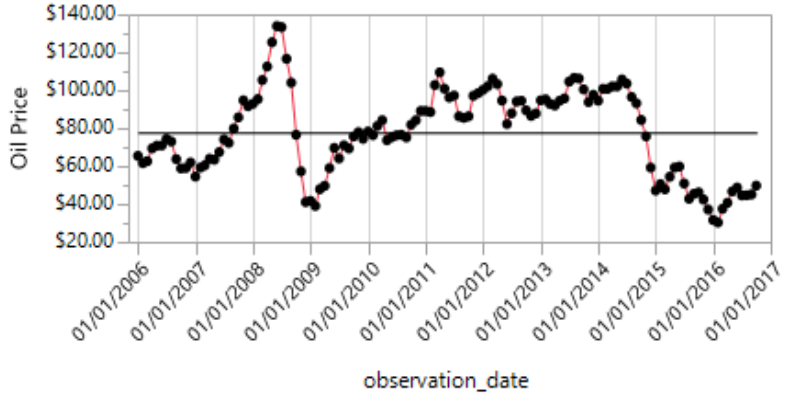
Transfer Function Analysis

Time Series Exchange Rate CA$ in US$



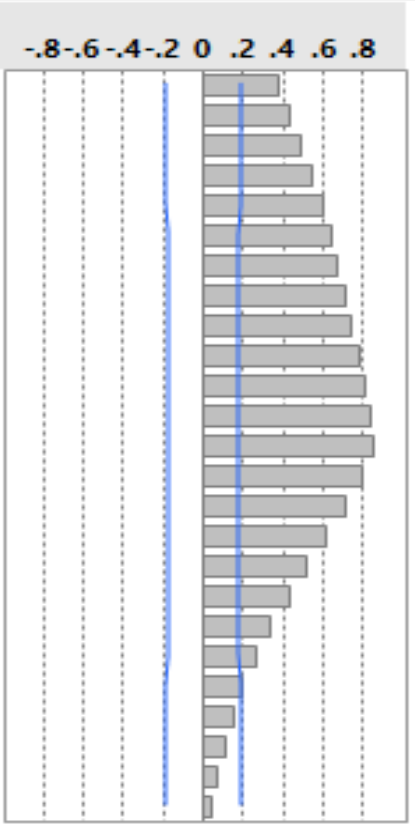
Input Time Series Panel

Input Series: Oil Price



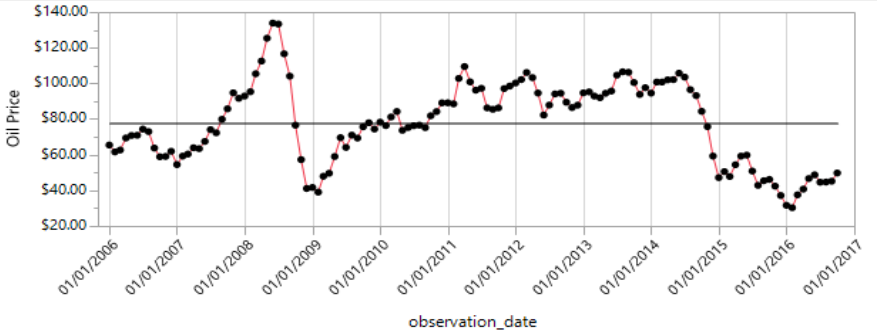
Cross Correlation Plots: Output Series - Exchange Rate CA$ in US$

Lag	Corr vs. Oil Price
-12	0.3829
-11	0.4402
-10	0.4943
-9	0.5491
-8	0.5968
-7	0.6382
-6	0.6761
-5	0.7108
-4	0.7461
-3	0.7808
-2	0.8174
-1	0.8431
0	0.8489
1	0.7932
2	0.7151
3	0.6219
4	0.5243
5	0.4297
6	0.3419
7	0.2651
8	0.2023
9	0.1497
10	0.1075
11	0.0679
12	0.0350

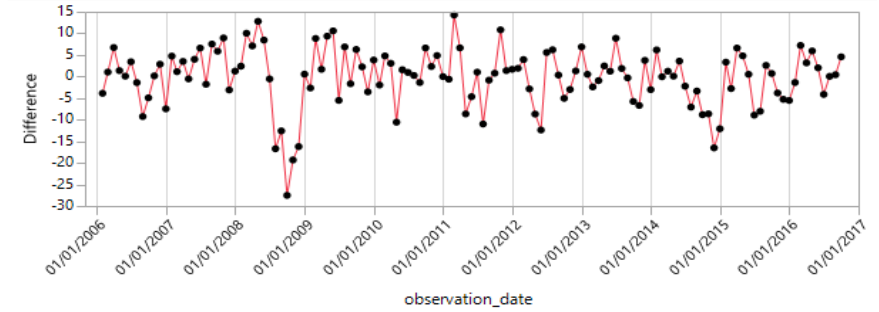


Modeling exchange rate as a function of oil price

Input Series: Oil Price



Difference: (1-B)^1



Lag	AutoCorr	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
0	1.0000		.	.	0	1.0000	
1	0.4388		25.4250	<.0001*	1	0.4388	
2	0.2466		33.5199	<.0001*	2	0.0670	
3	0.0183		33.5650	<.0001*	3	-0.1394	
4	-0.0471		33.8647	<.0001*	4	-0.0307	
5	-0.1635		37.5097	<.0001*	5	-0.1291	
6	-0.2454		45.7835	<.0001*	6	-0.1549	
7	-0.1681		49.6985	<.0001*	7	0.0317	
8	-0.1056		51.2571	<.0001*	8	-0.0009	
9	-0.0914		52.4328	<.0001*	9	-0.0820	
10	-0.0613		52.9657	<.0001*	10	-0.0248	
11	0.0243		53.0503	<.0001*	11	0.0463	
12	0.0767		53.9006	<.0001*	12	0.0163	

Model: ARI(1, 1)

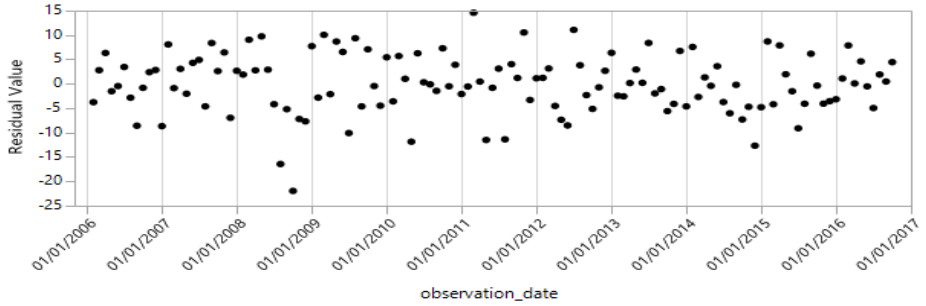
Model Summary

DF	127	Stable	Yes
Sum of Squared Errors	4644.51826	Invertible	Yes
Variance Estimate	36.5710099		
Standard Deviation	6.04739696		
Akaike's 'A' Information Criterion	832.587706		
Schwarz's Bayesian Criterion	838.307331		
RSquare	0.93020252		
RSquare Adj	0.92965293		
MAPE	6.64299312		
MAE	4.71321724		
-2LogLikelihood	828.587706		

Parameter Estimates

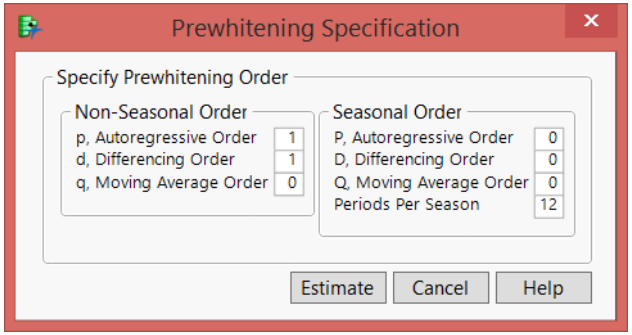
Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate
AR1	1	0.4382086	0.0788537	5.56	<.0001*	-0.0651155
Intercept	0	-0.1159069	0.8876074	-0.13	0.8963	

Residuals



Lag	AutoCorr	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
0	1.0000		.	.	0	1.0000	
1	-0.0268		0.0945	0.7585	1	-0.0268	
2	0.1163		1.8958	0.3876	2	0.1157	
3	-0.0804		2.7631	0.4296	3	-0.0756	
4	0.0071		2.7699	0.5970	4	-0.0096	
5	-0.0830		3.7086	0.5921	5	-0.0664	
6	-0.1804		8.1780	0.2254	6	-0.1932	
7	-0.0577		8.6384	0.2797	7	-0.0528	
8	-0.0116		8.6572	0.3720	8	0.0153	
9	-0.0438		8.9276	0.4440	9	-0.0649	
10	-0.0555		9.3647	0.4979	10	-0.0766	
11	0.0260		9.4613	0.5794	11	0.0059	
12	0.1170		11.4400	0.4916	12	0.0848	
13	-0.0498		11.8016	0.5440	13	-0.0831	
14	-0.0461		12.1139	0.5972	14	-0.0869	
15	-0.0716		12.8748	0.6120	15	-0.0806	

Modeling exchange rate as a function of oil price



Prewhitening Plot (1,1,0)(0,0,0)12 : Output - Exchange Rate CA$ in US$ vs. Input - Oil Price

Lag	Prewhiten Corr
-25	0.1264
-24	-0.0900
-23	0.1644
-22	0.0482
-21	0.0181
-20	-0.0629
-19	0.0978
-18	0.0112
-17	-0.0746
-16	-0.0884
-15	-0.0845
-14	-0.0858
-13	-0.0311
-12	0.0489
-11	0.0925
-10	-0.0395
-9	0.0636
-8	-0.0108
-7	-0.0322
-6	0.1041
-5	-0.0760
-4	0.0585
-3	-0.1323
-2	0.1070
-1	0.0064
0	0.5725
1	-0.0137
2	0.1149
3	-0.0200
4	0.0494
5	-0.0638
6	-0.1448
7	-0.0962
8	-0.1187
9	-0.1047
10	0.1263
11	-0.0891
12	0.0434
13	-0.0513
14	0.0163
15	0.0318
16	-0.1012
17	0.0250
18	-0.0039
19	0.0485
20	-0.1374
21	0.2164
22	0.0582
23	0.0014
24	-0.1375
25	0.1271

Modeling exchange rate as a function of oil price

Transfer Function Model (2)

Model Summary

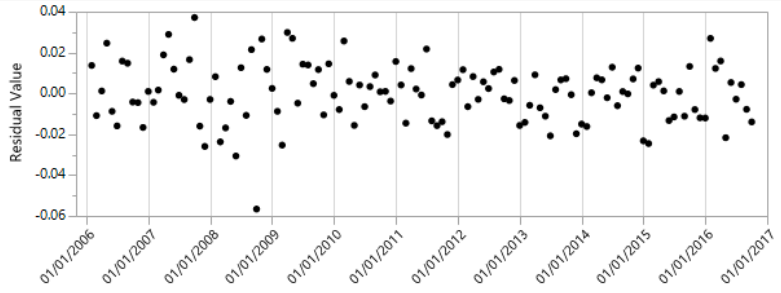
DF	126
Sum of Squared Errors	0.026542
Variance Estimate	0.00021065
Standard Deviation	0.01451381
Akaike's 'A' Information Criterion	-722.90994
Schwarz's Bayesian Criterion	-714.33051
RSquare	0.43332625
RSquare Adj	0.42433143
MAPE	1.25019787
MAE	0.01125327
-2LogLikelihood	-728.90994

Parameter Estimates

Variable	Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
Oil Price	Num0,0	0	0	0.0017845	0.0002033	8.78	<.0001*
Exchange Rate CA$ in US$	MA1,1	1	1	-0.2493661	0.0897089	-2.78	0.0063*
	Intercept	0	0	-0.0006290	0.0015819	-0.40	0.6916

$$(1 - B) * \text{Exchange Rate CA\$ in US\$}_t = -0.0006 + 0.0018 * (1 - B) * \text{Oil Price}_t + (1 + 0.2494 * B) * e_t$$

Residuals



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	-0.0191	0.0481	0.8264	1	-0.0191
2	-0.0990	1.3519	0.5087	2	-0.0994
3	-0.0755	2.1159	0.5487	3	-0.0803
4	0.1225	4.1461	0.3866	4	0.1107
5	0.1369	6.7000	0.2439	5	0.1308
6	-0.0322	6.8421	0.3357	6	-0.0093
7	-0.0807	7.7444	0.3557	7	-0.0437
8	-0.0334	7.9000	0.4433	8	-0.0376
9	-0.0918	9.0862	0.4294	9	-0.1448
10	-0.0339	9.2489	0.5087	10	-0.0736
11	0.0514	9.6269	0.5642	11	0.0481
12	-0.1704	13.8220	0.3122	12	-0.1755
13	-0.0088	13.8332	0.3857	13	0.0172
14	-0.0422	14.0946	0.4427	14	-0.0258
15	0.0725	14.8731	0.4606	15	0.0422
16	0.0545	15.3181	0.5015	16	0.0706
17	-0.1438	18.4383	0.3617	17	-0.1133
18	-0.0338	18.6118	0.4161	18	-0.0423
19	0.0840	19.6969	0.4130	19	0.0370
20	0.0597	20.2496	0.4424	20	0.0060
21	0.0318	20.4078	0.4956	21	0.0226
22	-0.0598	20.9728	0.5224	22	-0.0168
23	0.0564	21.4807	0.5518	23	0.0759
24	0.0734	22.3489	0.5584	24	0.0223
25	-0.0710	23.1674	0.5678	25	-0.0697

Modeling exchange rate as a function of oil price

Transfer Function Model (2)

Model Summary

DF	126
Sum of Squared Errors	0.026542
Variance Estimate	0.00021065
Standard Deviation	0.01451381
Akaike's 'A' Information Criterion	-722.90994
Schwarz's Bayesian Criterion	-714.33051
RSquare	0.43332625
RSquare Adj	0.42433143
MAPE	1.25019787
MAE	0.01125327
-2LogLikelihood	-728.90994

Parameter Estimates

Variable	Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
Oil Price	Num0,0	0	0	0.0017845	0.0002033	8.78	<.0001*
Exchange Rate CA$ in US$	MA1,1	1	1	-0.2493661	0.0897089	-2.78	0.0063*
	Intercept	0	0	-0.0006290	0.0015819	-0.40	0.6916

$$(1 - B) * \text{Exchange Rate CA\$ in US\$}_t = -0.0006 + 0.0018 * (1 - B) * \text{Oil Price}_t + (1 + 0.2494 * B) * e_t$$

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.

MS (Analytics) Class of 2017

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.

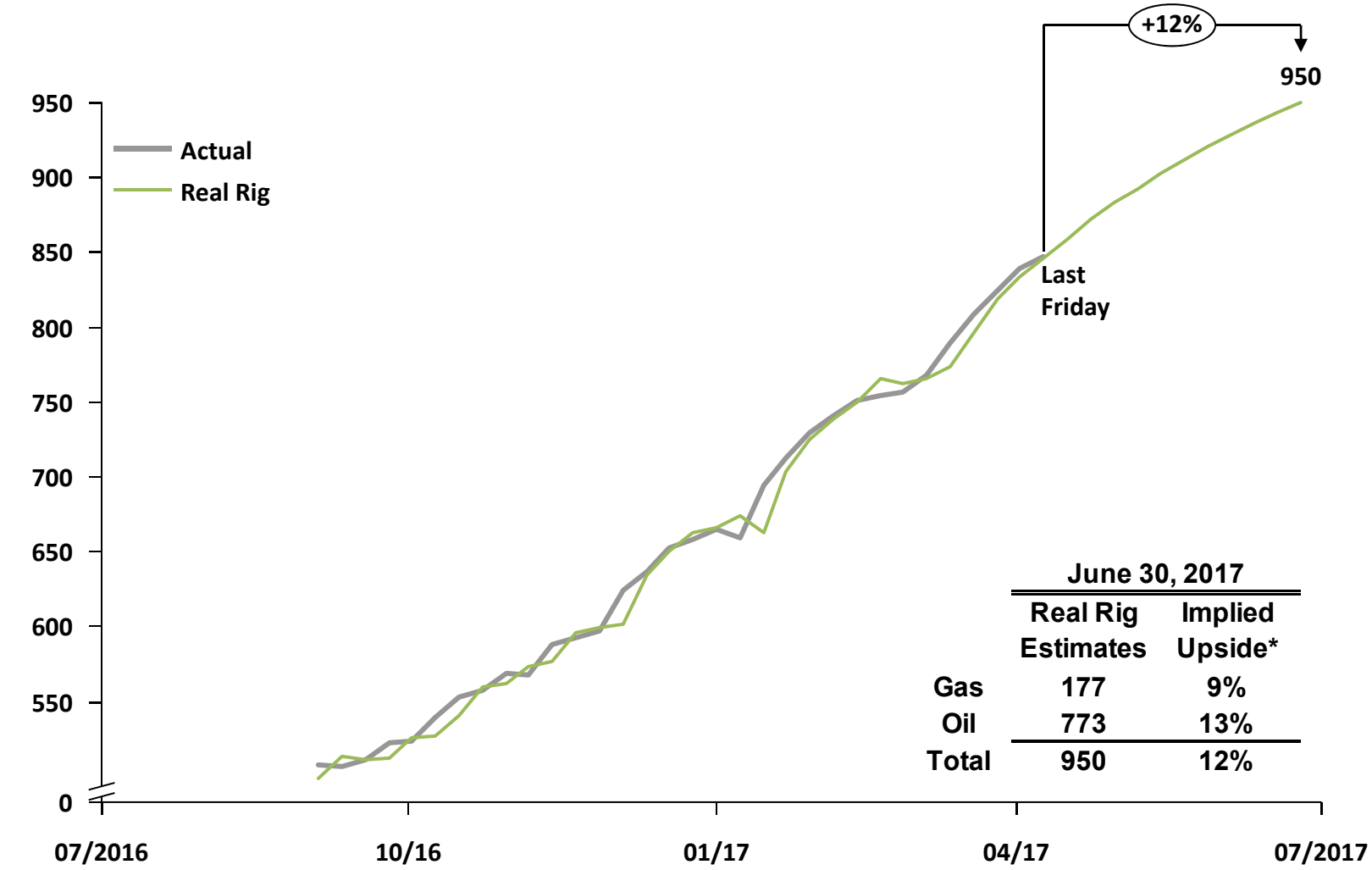
- Analysis was performed by Real Rig a start up company that grew out of the Texas A&M University Analytics Program.

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

- Transfer Function Model:
  $\text{Rig Count}_{(t)} = \alpha * \text{RigCount}_{(t-n)} + \beta * \text{InputX}_{(t-q)} + \omega * \text{InputY}_{(t-w)} \ldots + \varepsilon$

Email : info@realrig.com

Real Rig U.S. Rig Forecast

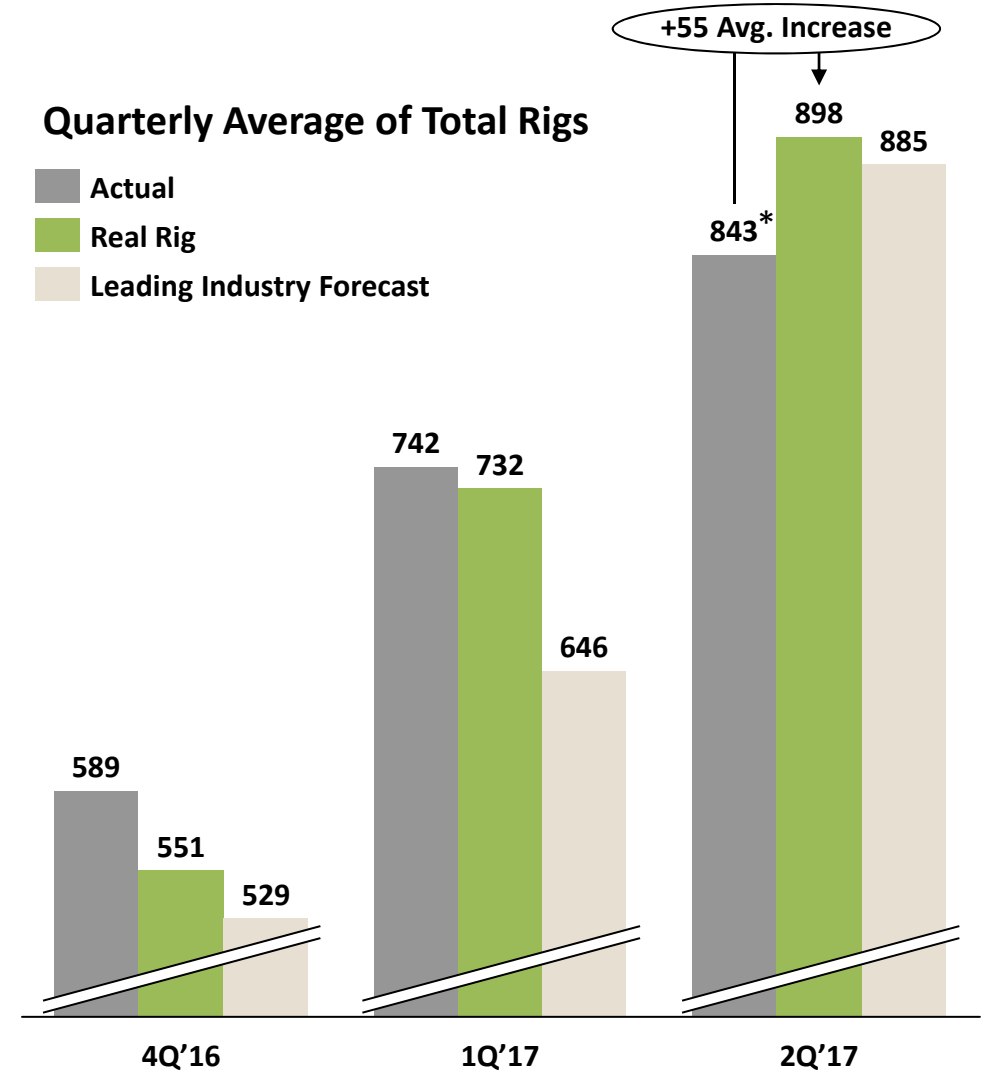


June 30, 2017	Real Rig Estimates	Implied Upside*
Gas	177	9%
Oil	773	13%
Total	950	12%

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

Email : info@realrig.com

Comparison to Industry



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

Outperforming leading industry forecast

Average Total Rigs

Quarter	Actual	Real Rig	Industry
4Q'16	589	551	529
1Q'17	742	732	646
2Q'17	843*	898	885

Percent Error

Quarter	Real Rig	Industry	Decreased Error
4Q'16	-6%	-10%	37%
1Q'17	-1%	-13%	90%

Email : info@realrig.com

Questions