Data Mining for Business (BUDT733)

Factors Driving Reproductive Health Supply Costs in Developing Countries

Team # 3

Jeffrey Chia-Chieh Hong, John Geraghty, Joseph Lee, Khalil Qasimi & Emma Spight

Fall 2009

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	**Signature**
Contact Author	Jeffrey Chia-Chieh Hong	
	John Geraghty	
	Joseph Lee	
	Khalil Qasimi	
	Emma Spight	

Executive Summary

In the last 40 years tremendous improvements have been made in the reproductive health of men and women in low and middle income countries. In most of the developing world, modern contraceptive use continues to increase with some estimates suggesting use will increase by 21% in the next 10 years. Key drivers include an increasing number of women of reproductive age and the desire of more people to plan their families. Our team obtained a large dataset of contraceptive supply data from John Snow Inc, a leader in Global Health Consulting. The data included factors such as quantity, cost, shipment data, country and manufacturer. Our goal was to understand which of the variables in the data were driving the cost of contraceptive supplies.

After analyzing the data using visualization techniques including scatter plots and bar graphs, we found it necessary to create a new Y variable called "Value per Couple's Year Protection (CYP)" (up to this stage we had considered unit cost as the Y variable).CYP is an industry standard calculation and enabled better comparison of the cost across the different contraceptives. Following this and further visualization, we created models using correlation, regression trees and multiple linear regression. From this analysis we were able to identify several key factors driving the supply costs.

The type of contraceptive procured is a key driver of cost.
Following visualizations and creating the couple year protection variable, it was apparent the different methods have very different cost structures. Appendix A shows the ranking of cost per CYP for each contraceptive. The regression trees and linear regression also support this finding. Before different contraceptives are procured, the procurer should review the different costs.

The quantity of contraceptive procured makes a difference in the cost per couple year protection.
This finding is apparent from the scatter plot in Appendix B. The regression trees and linear regression also support this finding. Quantity is an important measure for all contraceptives.

The manufacturer plays a key role in the cost per CYP.
Scatter pots revealed some measure of impact of the manufacturer. Using Regression trees solidified this finding as manufacturer was frequently at the top of the tree (Appendix C). Exhibit A on page 5 provides a ranking of the cost per CYP, low to high, for each manufacturer by contraceptive method.

Depending on the contraceptive, the procuring organization or the region drive the cost per CYP.
This finding became evident when regression trees were run for each method. As Appendix D highlights, region drives the cost for male condoms, IUD's, VTF and Implants. Procuring organization drives the cost for female condoms, orals combined and orals progestin only. Funding source impacts the cost for injectables and Standard day's method. However, due to the very high correlation (0.85) between procuring organization and funding organization it is reasonable to use procuring organization as a driver for these two contraceptives (See appendix E for correlation matrix).

Technical Description

Purpose
Our goal was to use explanatory data mining techniques to gain a better understanding of the variables driving the cost of reproductive health supplies. Our results will be used to guide future procurement decisions.

Data Source and Clean-up
Our original data consisted of 15,000 rows and 11 variables. We spent time exploring the data with Scatter Plots and summary statistics. Using this information and domain expertise, we dropped several variables leaving us with the seven listed below. We also removed any data before 2004 as a major procurer only started supplying data in 2004. All of our input variables, except for Quantity, are categorical. We performed several data preprocessing tasks including recoding and transforming the categorical variables to dummies. Using box plots we found several outliers in the data. Due to the quantity of data available, we decided removing these would have minimal impact on the analysis.

Variable Name	Description
Year	This is the year relating to shipment date. Date range is from 2004 to 2009
Funding Source	This is the organization paying for the product. The data consists of 27 different names.
Procuring Organization	This is the organization responsible for procuring the product. This could be the same as the Funding Source - e.g. USAID. Or, it could be a separate entity. There are currently 41 different organizations identified.
Method	Each product procured has been categorized. This data is clean and standardized. There are 12 categories.
Quantity	This is the total quantity purchased.
Manufacturer	This is the manufacturer who the procuring organization procures the product from. The manufacturer ships the product to the country.
Country	This is the receiving country. This data is clean, there are 155 different countries listed.
Region	The region of the receiving country. There are 6 different regions listed

Data Exploration
The exploration of the data has allowed us to come up with several interesting factors:
- **Introduce new Y variable Cost per Couple's Year Protection.** We originally intended for the Y variable to be unit cost. However, to allow comparison between the costs of different contraceptives, we created a new Y variable called Value per couple year protection (CYP). This is a standard industry identifier. The Value per CYP is how much it costs for a couple to have a year's worth of contraceptive supplies. For example, for birth control a couple needs 12 packs to be protected for one year, the cost per packet is $10 so the Cost per CYP is $120.
- **There is significant differential in the cost of per CYP for each method.** Foam/Jelly Spermicide is the most expensive costing on average $375 per CYP. Although there is no data for 2009 so potentially procurement of this method has ceased. VFT/Suppository

Spermicide is the second most expensive costing $147 per CYP. The least expensive method is IUD's costing on average $0.21 per CYP. The second least expensive method if Injectables costing on average $4.20 per CYP. ***Appendix A*** details these findings.

- **The quantity of each method procured does not align with the cost of each method.** Condoms are the most procured items, yet they rank 5th in cost per CYP. IUD's are the least expensive per CYP yet they rank much lower in quantity purchased. Appendix B details these findings
- **Cost per CYP decreases as Quantity increases.** Using a Scatter Plot it was clear to see a decrease in cost per CYP as Quantity increased. (***Appendix B***)
- **Manufacturer may play a role in cost per CYP** A scatter plot revealed some grouping of CYP based on manufacturer.

Modeling and Methods
There were three data mining methods were sued to determine the effect of each variable on the "Value per CYP": Correlation Analysis, Regression Trees and Linear Regression.

Correlation Analysis
We created a correlation matrix (Appendix E) to determine the relationship between the variables. From the matrix, we observed the strongest correlation (0.85) between Funding Source and Procuring Organization. Two other relationships that caught our attention were between Manufacturing Organization and Funding Source, and Manufacturing Organization and Procuring Organization. The rest of the relationships were insignificant.

Regression Trees
After analyzing the relationship between the different variables, we decided to use the Regression Tree method to further enhance our understanding of the data. We created a Regression Tree for each contraceptive method to confirm whether they share common cost factors. As a result, we have observed that the three top variables that drove the "Value per CYP" were Quantity, Method, and Manufacturers (***Appendix C***). As appendix D highlights procuring organization or region also play a role however this depends on the method.

Linear Regression
Multiple Linear Regression was the final method that was used for modeling. As the Regression Trees highlighted the predictors Quantity, Method and Manufacturer as the driving forces of CYP, we wanted to validate this using a different method. We used the p-value significance method to eliminate any non-significant predictors, one by one. As a result, the final predictors selected were Procuring Organization, Quantity, Method and Manufacturer. The error for the model was low, although not all the dummy variables for Procuring Org and Manufacturer had p-values lower than 0.05. This is expected as not all the Procuring Organizations or Manufacturers are linked with all the Methods. The Multiple R-Square was low however we believe this is due to other unknown variables not in the data source. (***Appendix F***)

Conclusion

Based on our exploration and modeling of the data, we have come up with several recommendations.

1) Based on the wide variation of costs per CYP the procurement strategy should vary for each Method. Appendix A shows the ranking of cost per CYP for each contraceptive.

2) The quantity procured is a driver in reducing the cost per CYP. The scatter plots, regression trees and regression supported this finding. Where possible procuring organizations should combine their procurement requests.

3) The manufacturer plays an important role in the cost per CYP. The table below lists the manufacturers that should be focused on depending on the method. The cost per CYP is listed from cost low to high.

4) Lastly, depending on the method, the procuring organization or the region drives the cost per CYP. As Appendix D highlights, region drives the cost for male condoms, IUD's, VTF and Implants. Procuring organization drives the cost for female condoms, orals combined and orals progestin only. Funding source impacts the cost for injectables and Standard day's method. However, due to the very high correlation (0.85) between procuring organization and funding organization it is reasonable to use procuring organization as a driver for these two contraceptives (See appendix E for correlation matrix).

**Exhibit A: Manufacturer Ranking - Low to High Cost per CYP**

Method	Manufacturer
Male Condom	UNITED MED-HEALTH INTERNATIONAL CORP
	MERCATOR HEALTHCARE LIMITED
	Ansell/Suretex Limited
Female Condom	THE FEMALE HEALTH CO UK PLC
	UNFPA
Oral - Combined	Famy Care Ltd
	WYETH
	Bayer Corp
Oral - Progestin	WYETH
	N V ORGANON
	Bayer
Oral - Emergency	MEDIMPEX UK LTD
	Famy Care Ltd
	Gedeon Richter
IUD	SMB Corporation of India
	Contech Devices
	Pregna International
Injectables	UNFPA
	Pfizer
	N V ORGANON
VFT / Spermicide	INNOTECH INTERNATIONAL
	Eisai Co., Ltd.
	UNFPA
Implants	UNFPA
	Schering
	Bayer Corp

Appendix A: Cost of per CYP for each method

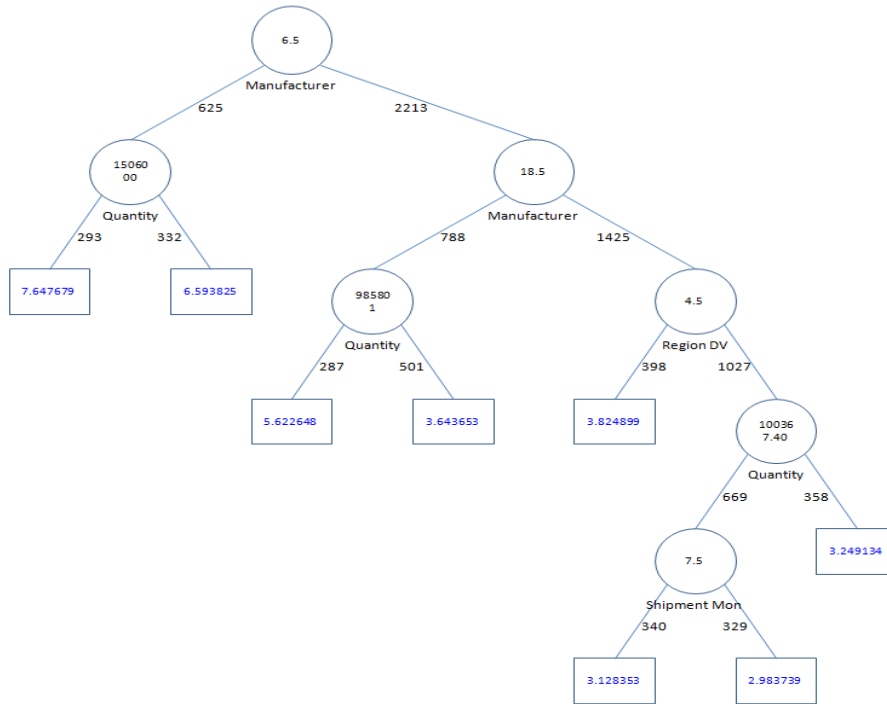| Rank	Method	Cost per CYP
1	IUDs	0.25
2	Standard Days Method	0.76
3	Injectables	4.20
4	Orals - Combined	4.99
5	Condoms - Male	5.04
6	Orals - Progestin Only	6.55
7	Implants	7.08
8	Orals - Emergency	7.17
9	Condoms - Female	95.85
10	VFT/Suppository\ Spermicide	147.43
11	Foam/Jelly Spermicide	382.08

Appendix B: Scatter Plot Value per CYP vs. Quantity
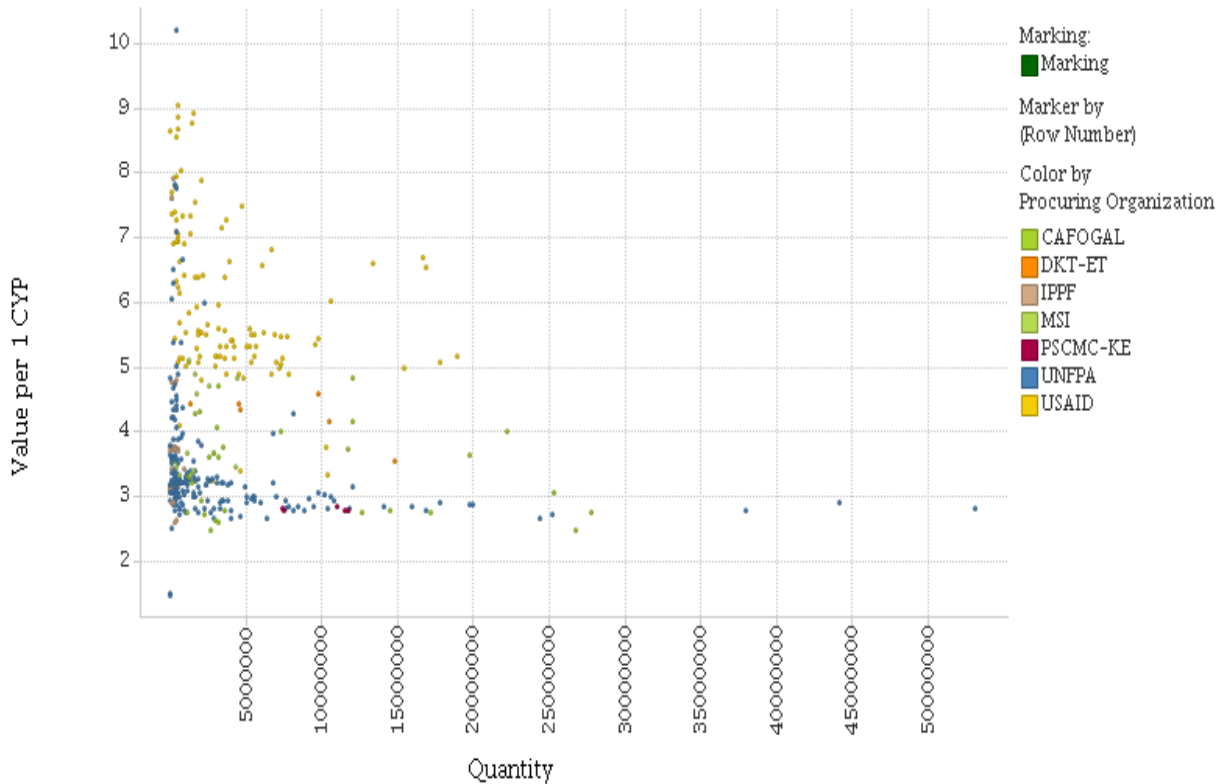
Appendix C: Regression

Tree Analysis

Appendix D: Table of Regression Trees, top 3 variables by method

	Shipment month	Funding Source	Procuring org	Quantity	Region	Country	Manufacturer
Male condom				X	X		X
Female condom	x		X	X			
Orals Combined			X	X			X
Oral Progestin Only			X	X			X
IUD's				X	X		X
Injectables		x				X	X
VTF	x			x	x		N/I
Standard days method							N/I
Implants				x	X	x	
Oral Emergency		x		x			x

Appendix E: Correlation Matrix

	Funding Source	Procuring Org	Method DV	Manufacturer DV	Country DV	Region DV	Quantity
Funding Source	1	0.854824041	-0.031800951	0.335852056	0.273101325	-0.028571985	-0.096385694
Procuring Org		1	0.102356179	0.371997587	0.279424018	0.007256085	-0.146162197
Method DV			1	0.084534452	0.008346844	0.018879781	-0.207686357
Manufacturer DV				1	0.149136983	-0.046703684	0.016275497
Country DV					1	-0.025617682	-0.206928671
Region DV						1	-0.033715085

Appendix F: Multiple Linear Regression Output

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3917.07471	13987.17969	0.77955782	1.24896E+12
Procuring Org Dummy 1	51436.77344	16513.20898	0.00194665	20733812736
Procuring Org Dummy 2	41670.53906	14048.37891	0.00316028	2.48349E+11
Procuring Org Dummy 3	-703.666077	26071.93359	0.97847903	1.21081E+11
Procuring Org Dummy 4	-41282.3945	55387.48828	0.45642051	25454684160
Procuring Org Dummy 5	22709.66406	14346.12598	0.11406288	58817777664
Procuring Org Dummy 7	-12147.5645	36394.27344	0.73868906	19268962304
Method Dummy 1	-39181.5586	13737.29688	0.00452295	47946600000
Method Dummy 2	-18897.873	28645.44336	0.50974196	1.40929E+11
Method Dummy 3	1789.234253	12233.85449	0.88378131	300311840
Method Dummy 4	-16422.8359	14073.98145	0.24381411	95069085696
Method Dummy 5	-120676.625	23404.40039	0.00000036	70731620352
Method Dummy 6	15709.61523	14343.08398	0.27392793	32313948160
Method Dummy 7	33452.40625	14388.96289	0.02048185	1.19773E+11
Quantity	0.01640807	0.00111715	0	2.86406E+12
Manufacturer Dummy 1	-12050.9141	18510.95117	0.51533806	35011469312
Manufacturer Dummy 2	-19688.1074	14995.42871	0.18981008	11519736832
Manufacturer Dummy 3	-43624.5977	19420.58789	0.02512361	23232600000
Manufacturer Dummy 4	-23020.0645	11532.34082	0.0464659	4.64706E+11
Manufacturer Dummy 6	27102.6543	21252.62695	0.20281371	1591175808
Manufacturer Dummy 7	151668.6094	21005.02734	0	4.76965E+11

Residual df	1385
Multiple R-squared	0.299261521
Std. Dev. estimate	104170.0703
Residual SS	1.50292E+13

Total sum of squared errors	RMS Error	Average Error
1.50292E+13	103023.4788	0.001541517