Headquarters U.S. Air Force

Integrity - Service - Excellence

A Methodology for Multivariate Regression on Large Datasets

AFCAA/FMS

Matt Pitlyk

SCEA June 2011

U.S. AIR FORCE

U.S. AIR FORCE

Overview

- Present motivation for ideas
- Give an approach for analysis of large datasets
- Present logical organization of work/outputs to make it understandable
- Show ways to automate procedures and collection/summary of outputs
- Not an Excel or VBA lesson. I'll should you what you can do, but not how to do it

U.S. AIR FORCE

Background

- Common practices
  - Make scatterplots, ANOVA tables, all over the same sheet trying to explore as many relationships as possible
  - Add trendlines for each possible model
  - Have to recalculate certain statistics which are not part of ANOVA tables over and over, e.g. CV
- Wish we could make scatterplotting and calculating all regressions and statistics easier and more organized

U.S. AIR FORCE

Background

- Was working with dummy (indicator/stratifying) variables to represent varying levels of categorical variables and wanted to be able to see those levels visually
  - E.g. Stratifying on F/A or B/C
- Needed to run many independent variable combinations
- Needed to run the same variable combinations on multiple datasets
  - Run same combinations on each subsystem
- Had to choose which variable combinations to run
- With 15 variables, that is over 32,000 combinations

U.S. AIR FORCE

Background

- Was asked to write a function that would produce loglinear statistics dynamically (update automatically)
- What if we had a scatterplot that could change variables more easily?
- What if we could automatic regression and collect the outputs in a consistent manner?

U.S. AIR FORCE

Procedure

- Assumptions **(based on metadata)**
  - **Identified sound independent variables based on engineering judgment**
  - **Identified exclusive relationships between IV's E.g. families of dummy variables**

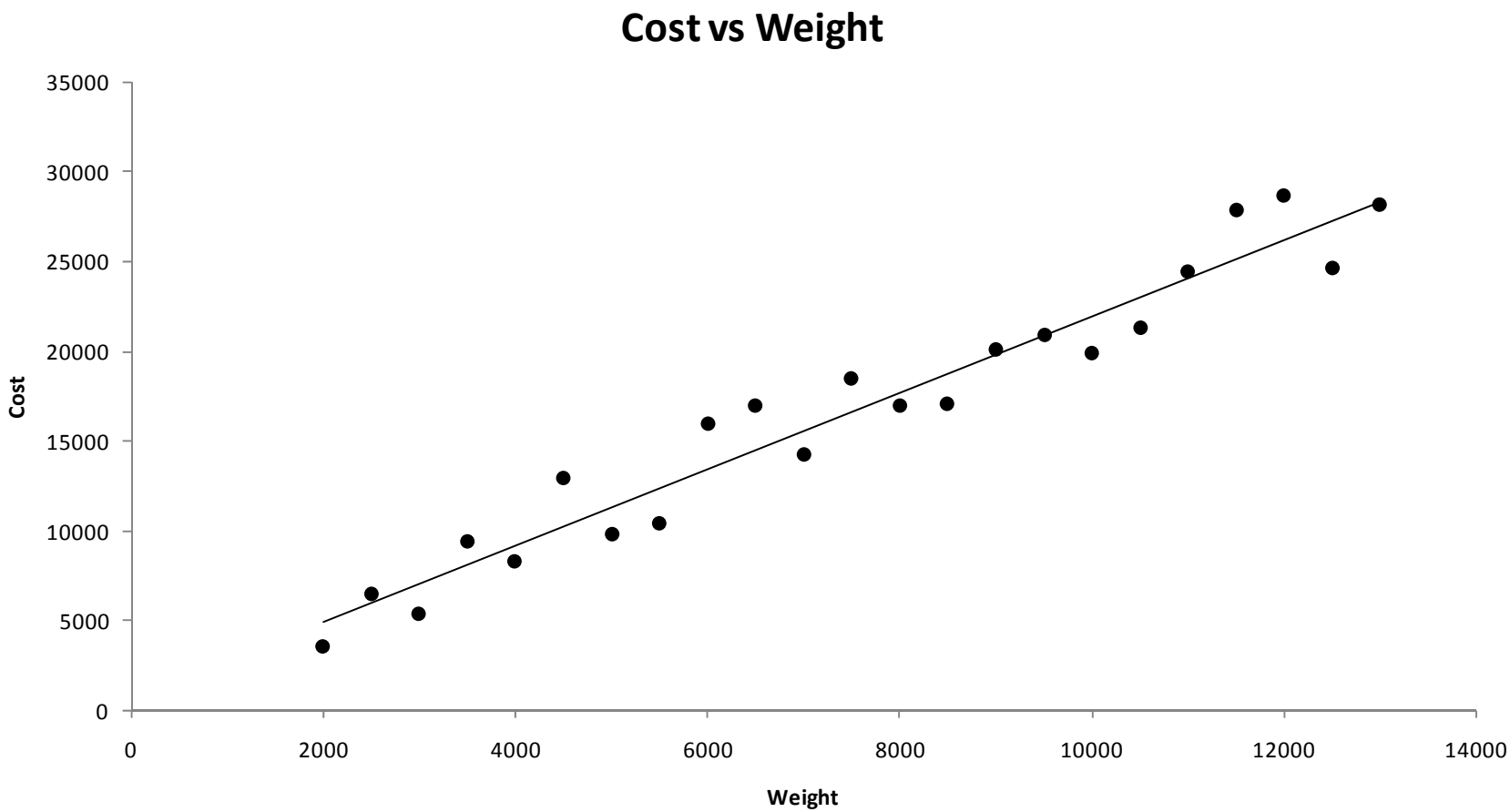
U.S. AIR FORCE

Procedure

- **Scatterplot variables that you expect to have good relationships (with engineering judgment) or have high correlation with the dependent variable**
  - **Use colors and sizes to display varying factor levels of categorical variables**
  - **Include data points names for quick inspections of anomalies**
- **Scatterplot categorical variable vs cost to get a rough idea of the mean and variance**

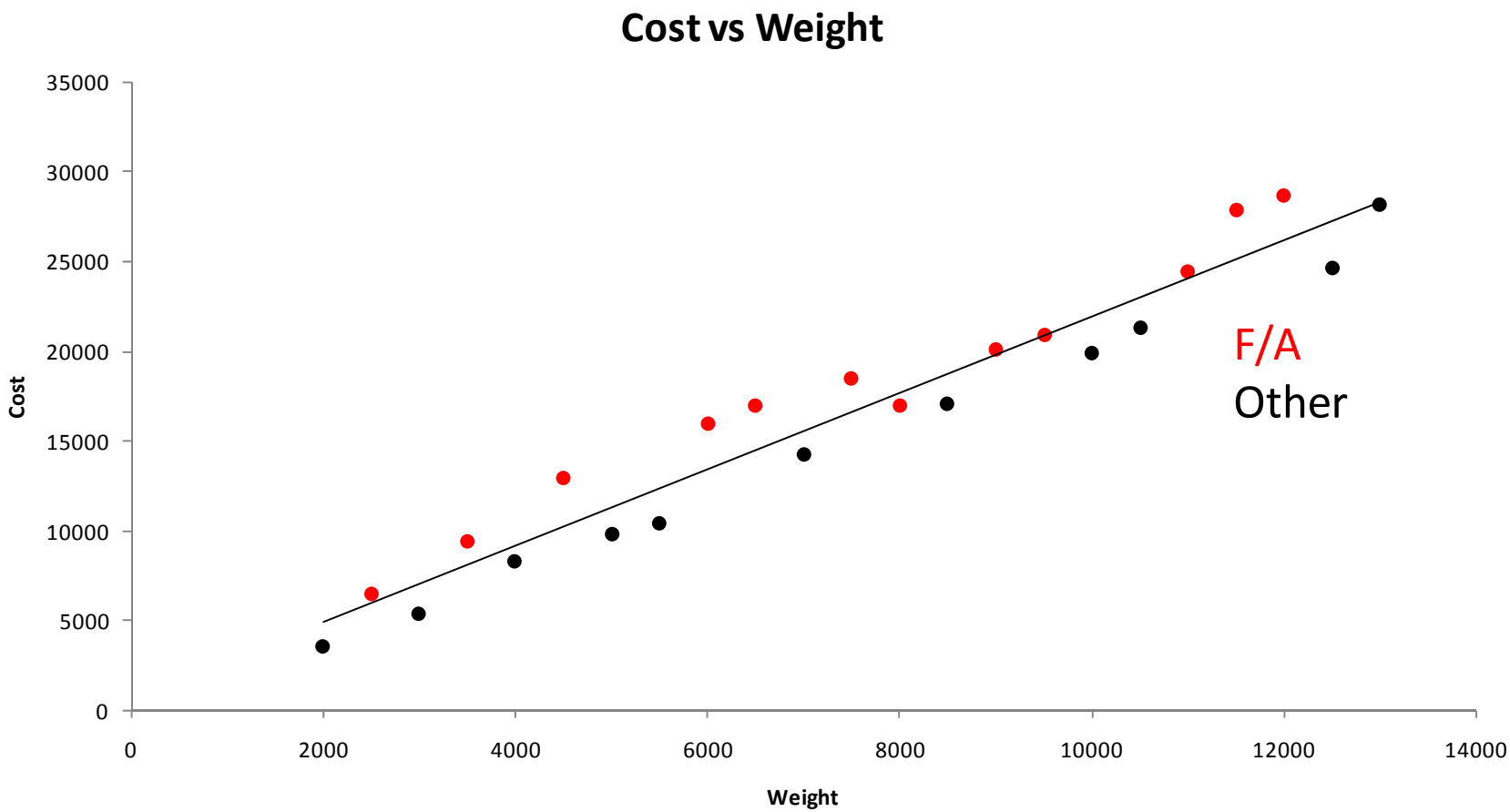
U.S. AIR FORCE

Procedure



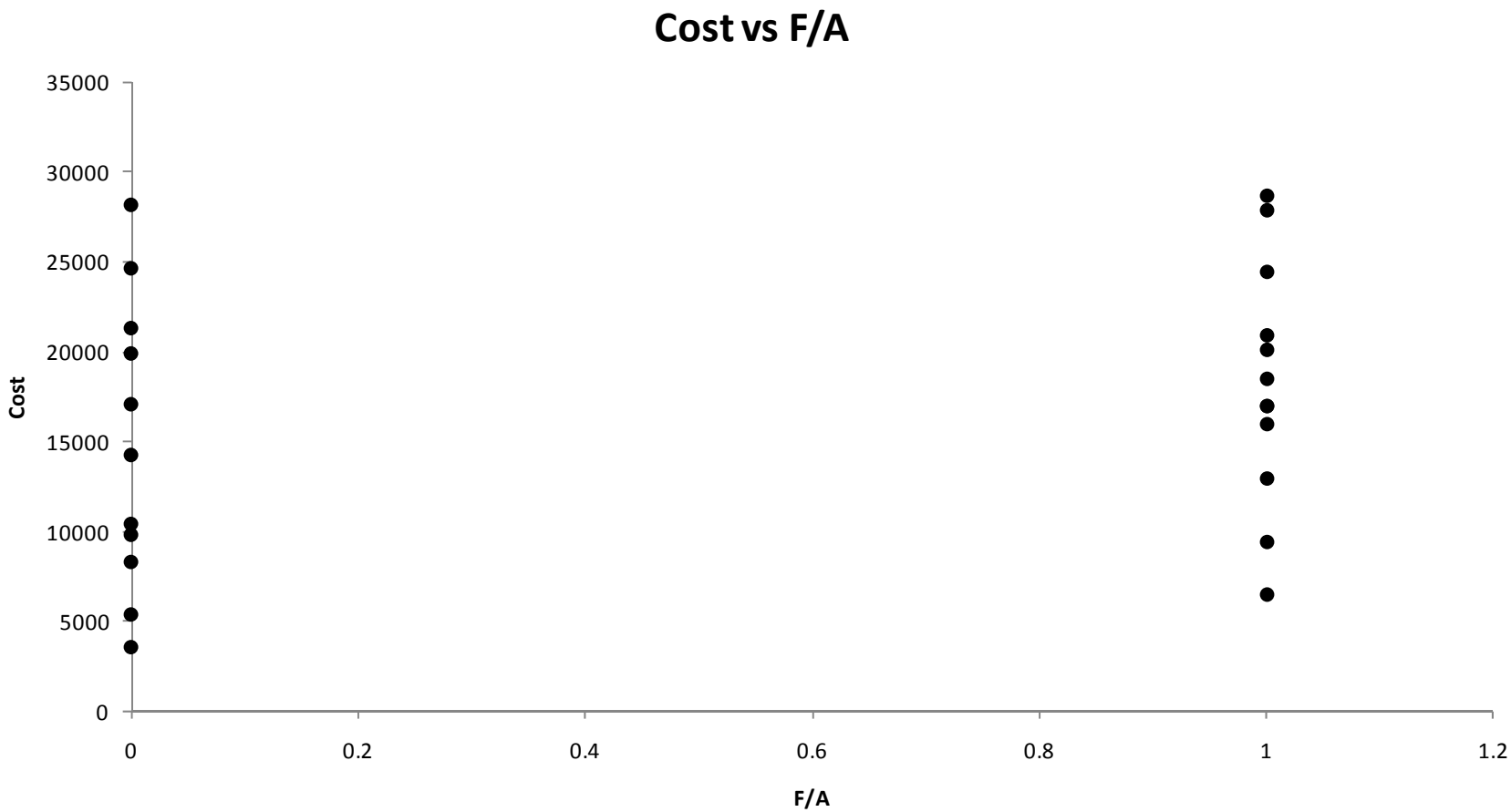
U.S. AIR FORCE

Procedure



U.S. AIR FORCE

Procedure



U.S. AIR FORCE

Procedure

Backwards Elimination: Take a collection of independent variables in which the biggest cost drivers do not have high correlation (Absolute Value of .7 or higher)

- e.g. Weight and power are big drivers for satellites, so do not start with both of them
- High correlation in smaller cost drivers is alright for now
- OK to have dummy variables within the same family
- Perform Backwards Elimination
  - Run regression
  - Remove least significant variable (highest p-value)
  - Rerun regression
  - Repeat until all remaining variables are significant (minimal equation)

U.S. AIR FORCE

Procedure

Backwards Elimination: An example.

Elimination value of .1

Estimate T1 cost of a satellite using Weight, EoL (Power), Design Life, Number of Payloads, and dummy variables for a GEO orbit and a MEO orbit. Weight and EoL are both highly correlated with cost and also highly correlated with each other, so we only start with Weight and perform a regression run which yields the p-values on the next slide:

U.S. AIR FORCE

Procedure

Wt	DL	Pay	GEO	MEO
.07	.12	.08	.22	.13

DL, GEO, and MEO all have p-values greater than our elimination level of .1, but we only remove the least significant variable, GEO. After running the regression again we have:

Wt	DL	Pay	MEO
.03	.15	.05	.12

Removing MEO has affected all the p-values and now DL is the least significant variable and is greater than our elimination level. We remove DL and run the regression again…

U.S. AIR FORCE

Procedure

Wt	Pay	MEO
.02	.06	.09

Removing DL affected all the p-values for the remaining variables, but now they are all less than our elimination value of .1 and so we have our minimal equation.

Now we must inspect it.

U.S. AIR FORCE

Procedure

- **Inspect your minimal equation for:**
  - **High correlation between remaining variables**
  - **2+ dummy variables from the same family**

- **Run combination equations**
  - **If Design Life and Number of Payloads are highly correlated, remove DL from the minimal equation and rerun regression, then remove Payloads from minimal and rerun**
  - **If the dummy for satellites with LEO orbits and the dummy for MEO orbits are both in your minimal equation, rerun regression with each one removed**
  - **Suggest Backward Elimination starting with these combination equations**

U.S. AIR FORCE

Procedure

If the minimal equation from our example contained the variables

Weight, DL, Pay, GEO, MEO

We would next run four more regressions with theses combinations:

Weight, DL, GEO

Weight, DL, MEO

Weight, Pay, GEO

Weight, Pay, MEO

Once all combination runs are complete, we start over with EoL instead of Weight as the biggest cost driver and repeat all the steps.

U.S. AIR FORCE

Caveats

- **The more regressions you run, the greater the chance of a false positive (Type I error)**
  - **Reserve data to test regression coefficients against**
  - **Reserving is difficult if you are trying to stratify your data**

- **Removing some variable might cause a previously removed variable to become significant, if it were placed into the regression again**
  - **If you found a relationship during scatterplotting, try replacing previously removed variables after getting a minimal equation**
  - **Cannot try every combination, some relationships are bound to be missed, but the number can be reduced with careful analysis**

U.S. AIR FORCE

Organization

- **Organization: Organization of inputs and outputs may be undervalued. Inputs, computations, and outputs are separated for models, why not during model development (regression analysis) too!**
  - **One run per sheet**
  - **Create a template for your output and STICK TO IT! Easier to read and to find what you are looking for - also makes automation easier**
  - **Create a summary sheet to display most important information for comparing potential CERs**

U.S. AIR FORCE

Automation

- Record a macro for running ANOVA and adding formulas for stats
- Create a template that is just waiting for the ANOVA output to appear in a specific spot
- Create a tool the does the regression and formatting for you

Automation

U.S. AIR FORCE

Wt_F.A

Regression Type: Linear

Equation: *Cost = 471.591 + Weight*1.998 + F/A*2599.136*

		Intercept	Weight	F/A
Output:	471.591	471.591	0	0

Alpha level:	0.1
Data points:	23

Variables

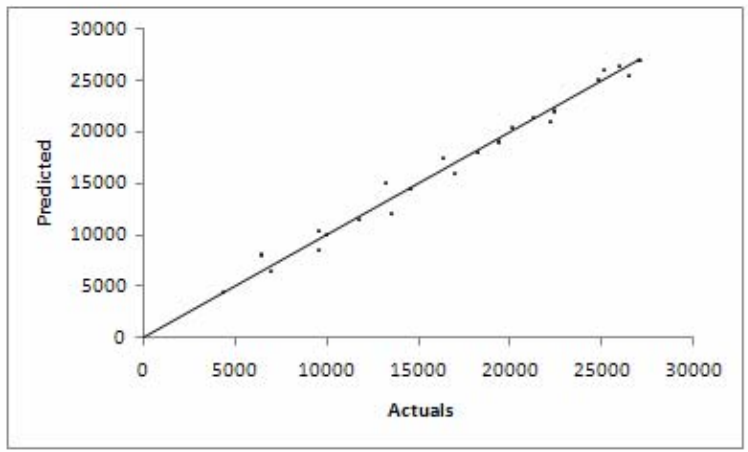
Name	Intercept	Weight	F/A
Coefficie	471.591	1.998	2599.136
SE	501.104	0.057	378.697
P-value	0.358	0.000	0.000

Regression Statistics

R^2	0.985	Regression DF	2	SSR	1063912748.006
Adj R^2	0.983	Residual DF	20	SSE	16435627.822
F-Stat	647.321	Total DF	22	SST	1080348375.829
F-Stat P-V	0.000			SEE	906.522
Pearson'	0.992			Y_Bar	16809.612
				CV	0.054

Red Flag if Abs >=	0.7
Yellow Flag if Abs >=	0.6

	Cost	Weight	F/A
Cost	1.0000	0.9741	0.2275
Weight	0.9741	1.0000	0.0394
F/A	0.2275	0.0394	1.0000



Automation

U.S. AIR FORCE

RAND | =INDIRECT(A4&"!I18")

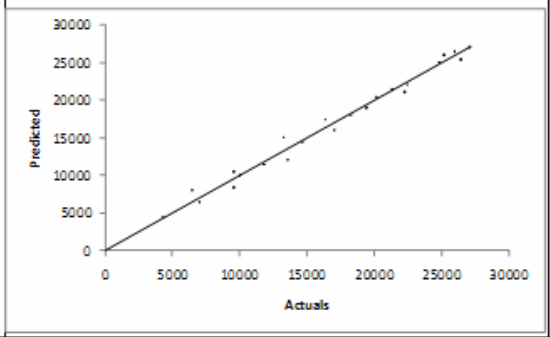
	Fit Statistics					Significance tests (p-values shown)			
Tab Name	CV (Unit)	SEE (Unit)	Adj R^2 (Unit)	SEE_log	Adj R^2 (Fit space)	Var_1	Var_2	Var_3	Reg (F-Stat)
CER 1	0.6562	5387.5298	0.4342			0.00000			0.00000
CER 2	0.6516	(A4&"!I18")	0.5263			0.00000			0.00000
CER 3	0.7692	5373.9748	0.4751	0.5203	0.4521	0.00000			0.00000
CER 4	0.7875	5055.5934	0.5367	0.5018	0.4784	0.00000	0.26600		0.00000
CER 5	0.6657	5328.2824	0.3670	0.4490	0.5730	0.00000		0.09620	0.00000
CER 6	0.747941535	4977.146586	0.494712612	0.6048	0.5535	0.00000	0.15910	0.08760	0.00000
Best	0.717502686	5305.482027	0.495237482	0.5359	0.5131	0.00000		0.06080	0.00000

Summary | CER 1 | CER 2 | CER 3 | CER 4 | CER 5 | CER 6 | Best

Automation

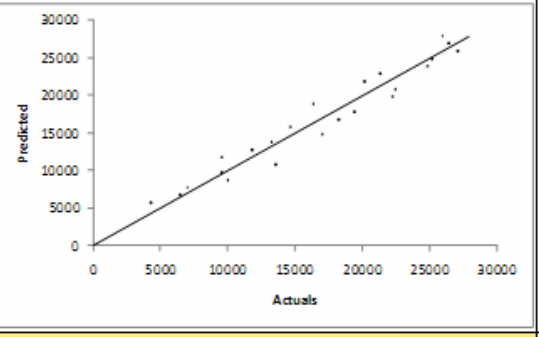
U.S. AIR FORCE

Run #	Name	DV	Variables	Coefficient	P-Value	Error DF	CV (Unit)	SEE (Unit)	Pearson's Co R^2 (Unit)	Adj R^2 (Unit)	DP	EQN
1	Wt_F/A	Cost	Intercept	471.5910347	0.357880925	20	0.0539288	906.5215889	0.992364214	0.984786733	0.983265407	23	Cost = 471.591 + Weight*1.998 + F/A*2599.136
			Weight	1.997593369	2.00539E-19							
			F/A	2599.136046	1.14156E-06							



	Cost	Weight	F/A
Cost	1.0000	0.9741	0.2275
Weight	0.9741	1.0000	0.0394
F/A	0.2275	0.0394	1.0000

Run #	Name	DV	Variables	Coefficient	P-Value	Error DF	CV (Unit)	SEE (Unit)	Pearson's Co R^2 (Unit)	Adj R^2 (Unit)	DP	EQN
2	Wt	Cost	Intercept	1712.087784	0.053138284	21	0.0964031	1620.49804	0.974143257	0.948955084	0.946524374	23	Cost = 1712.088 + Weight*2.013
			Weight	2.013003266	4.78113E-15							



	Cost	Weight
Cost	1.0000	0.9741
Weight	0.9741	1.0000

Run #	Name	DV	Variables	Coefficient	P-Value	Error DF	CV (Unit)	SEE (Unit)	Pearson's Co R^2 (Unit)	Adj R^2 (Unit)	DP	EQN
3	Wt_Payload	Cost	Intercept	1619.035287	0.286097088	20	0.098769	1660.268426	0.97415108	0.948970326	0.943867359	23	Cost = 1619.035 + Weight*2.014 + Payload Cap*0.002
			Weight	2.013867601	2.38893E-14							
			Payload Cap	0.002301861	0.939160229							

	Cost	Weight	Payload Cap
Cost	1.0000	0.9741	-0.0999

U.S. AIR FORCE

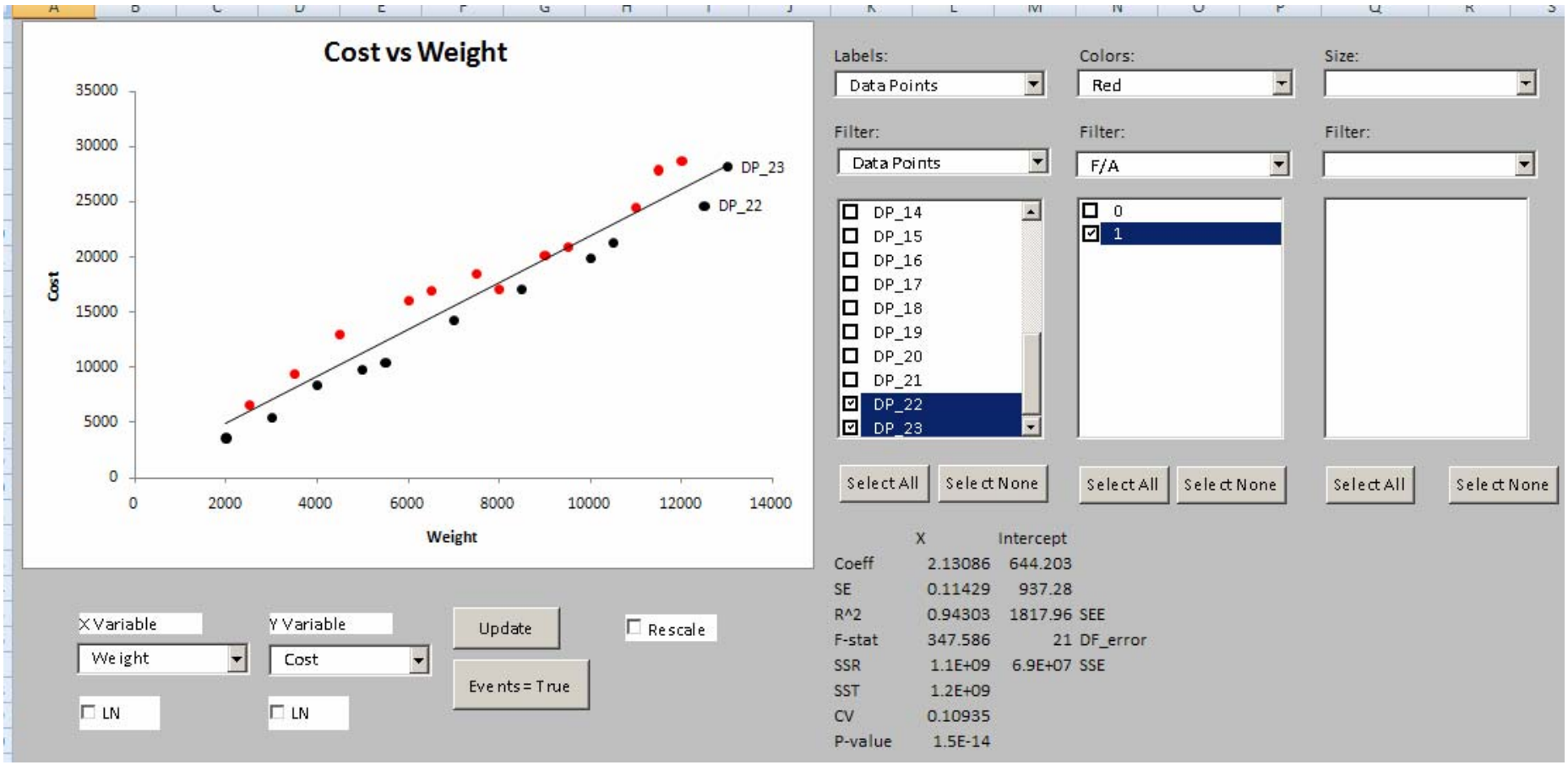
Automation

Run #	Name	DV	Variables	Coefficient	P-Value	Error DF	CV (Unit)	SEE (Unit)	Pearson's Co R^2 (Unit)	Adj R^2 (Unit)	DP	EQN
1	Wt_F,A	Cost	Intercept	471.5910347	0.357880925	20	0.0539288	906.5215889	0.992364214	0.984786733	0.983265407	23	Cost = 471.591 + Weight^1.998 + F/A^2599.136
2	Wt	Cost	Intercept	1712.087784	0.053138284	21	0.0964031	1620.49804	0.974143257	0.948955084	0.946524374	23	Cost = 1712.088 + Weight^2.013
3	Wt_Payload	Cost	Intercept	1619.035287	0.286097088	20	0.098769	1660.268426	0.97415108	0.948970326	0.943867359	23	Cost = 1619.035 + Weight^2.014 + Payload Cap^0.002
4	Wt_Payload_F,	Cost	Intercept	1185.956377	0.156232307	19	0.053551	900.171125	0.99284902	0.985749177	0.983499048	23	Cost = 1185.956 + Weight^1.99 + Payload Cap^-0.019 + F/A^2677.884
5	Payload_F,A	Cost	Intercept	18185.01777	0.001710364	20	0.4211974	7080.164919	0.268308472	0.071989436	-0.02081162	23	Cost = 18185.018 + Payload Cap^-0.085 + F/A^3471.045
6	Payload	Cost	Intercept	19011.97135	0.001065982	21	0.4245579	7136.654509	0.099890267	0.009978065	-0.037165836	23	Cost = 19011.971 + Payload Cap^-0.059

- **Use conditional formatting to highlight threshholds (p-values and correlation) and make comparisons easier**
- **Collapse rows to make comparisons easier**
- **Create links to the full regression sheets**

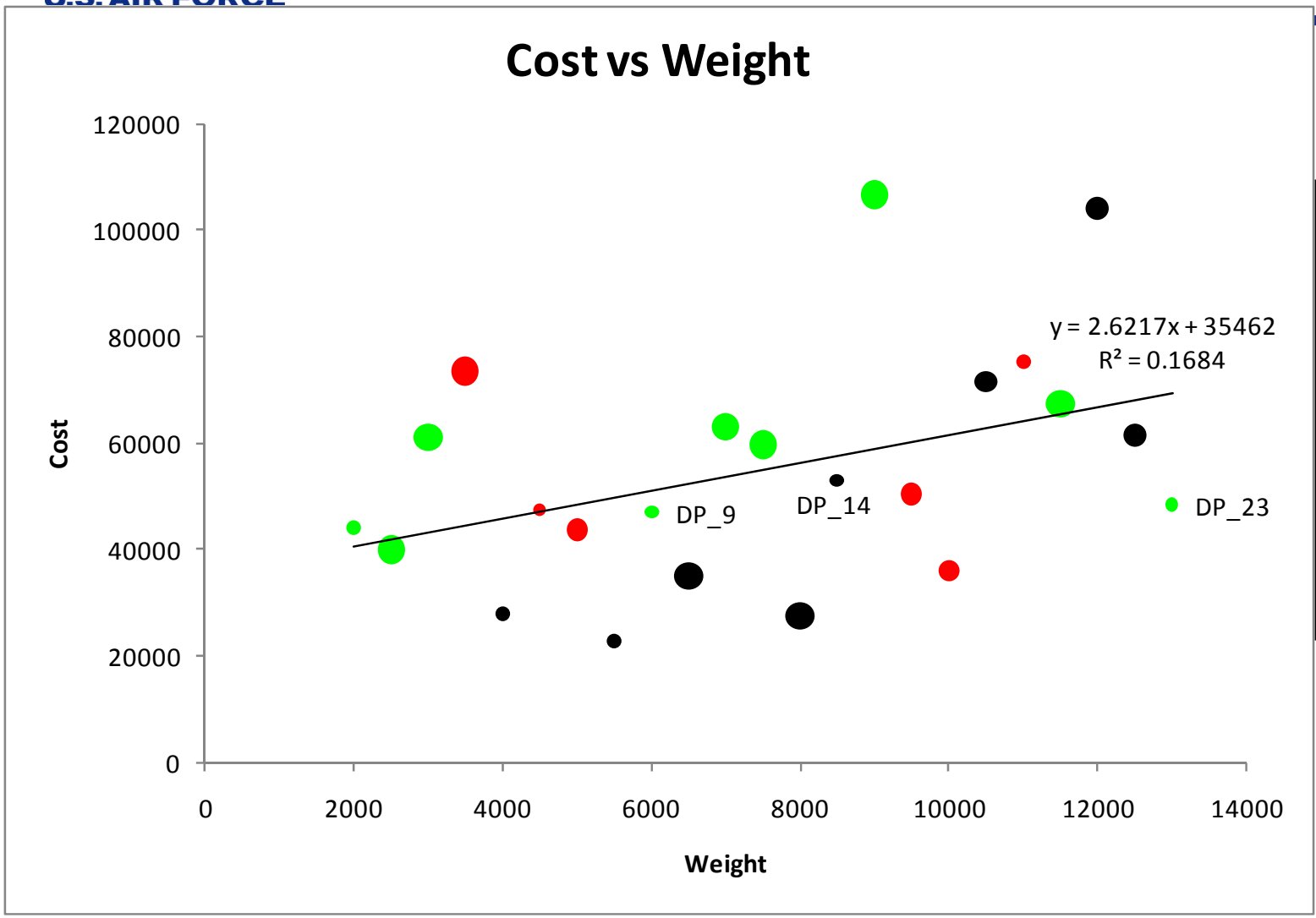
Automation

U.S. AIR FORCE



U.S. AIR FORCE

Automation



Orbits
Red:GEO
Green:MEO
Black:Other

DL (years)
Small : 1-4
Med : 5-8
Large : 9+

Automation

U.S. AIR FORCE