NORTHROP GRUMMAN

DEFINING THE FUTURE

Multicollinearity: Don't Throw Out The Baby (Independent Variable) with the Bath Water

Jim Byrd

Northrop Grumman Corporation

Purpose

NORTHROP GRUMMAN

- The purpose of this paper is to stimulate thought and suggest alternative solutions for analysts faced with the issue of multicollinearity

Agenda

NORTHROP GRUMMAN

- What Is Multicollinearity?
- How Is It Detected?
- What Are The Side Effects?
- What Are The Standard Solutions?
- What Are Some Not So Standard Solutions?
- Summary

What Is Multicollinearity?

NORTHROP GRUMMAN

- Multicollinearity occurs when two or more of the independent variables used in regression analysis are highly correlated with each other which is another way of saying that they move together
  - This may occur because they are both measuring the same thing though it may not be obvious
  - This may occur because they have A common/shared relationship
    - For example, most cost and price indices trend with time and therefore exhibit A statistical relationship with each other through their shared relationship with time
    - Another common example in production programs the number of cumulative units produced and the quantity produced per time period increase together at the beginning of the program
  - This may be A result of randomness
  - Note that the moving together may be in opposite directions

How Is Multicollinearity Detected?

NORTHROP GRUMMAN

- The easiest way to detect multicollinearity is by calculating the correlation between the independent variables

INDEX	*NON-METALLIC*	*INDUSTRIAL COMMODITIES*	*METALS & METAL PROD*	*TOTAL COMP AIRCRAFT MFG*
NON-METALLIC	1.00000			
INDUSTRIAL COMMODITIES	0.94116	1.00000		
METALS & METAL PROD	0.95332	0.91005	1.00000	
TOTAL COMP AIRCRAFT MFG	0.87113	0.95444	0.81333	1.00000

How Is Multicollinearity Detected?

NORTHROP GRUMMAN

- The easiest way to detect multicollinearity is by calculating the correlation between the independent variables

Notional Aircraft Program		
	QTY	*MIDPOINT*
QTY	1.00000	
MIDPOINT	0.78886	1.00000

What Are The Side Effects?

NORTHROP GRUMMAN

- The results from regression analysis may exhibit:
  - Coefficient(s) with the “wrong” sign
  - Coefficient(s) that are not significantly different from zero – the student t-test

What Are The Side Effects?

NORTHROP GRUMMAN

- The easiest way to detect multicollinearity is by calculating the correlation between the independent variables

INDEX	*NON-METALLIC*	*INDUSTRIAL COMMODITIES*	*METALS & METAL PROD*	*TOTAL COMP AIRCRAFT MFG*
NON-METALLIC	1.00000			
INDUSTRIAL COMMODITIES	0.94116	1.00000		
METALS & METAL PROD	0.95332	0.91005	1.00000	
TOTAL COMP AIRCRAFT MFG	0.87113	0.95444	0.81333	1.00000

What Are The Side Effects?

NORTHROP GRUMMAN

TOTAL COMP AIRCRAFT MFG

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.96384
R Square	0.92900
Adjusted R Square	0.92586
Standard Error	0.09470
Observations	72

ANOVA

	df	*SS*	*MS*	*F*
Regression	3	7.97846	2.65949	296.6
Residual	68	0.60981	0.00897	
Total	71	8.58827		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-1.22384	0.17638	-6.93847	0.00000
NON-METALLIC	0.10211	0.20174	0.50617	0.61438
INDUSTRIAL COMMODITIES	3.07644	0.24334	12.64266	0.00000
METALS & METAL PROD	-0.96881	0.29081	-3.33142	0.00140

What Are The Standard Solutions?

NORTHROP GRUMMAN

- Drop the offending variable(s) or replace it (them) with A more definitive variable(s)
- Gather more data
- Combine the collinear variables somehow into A new variable

Drop The Offending Variable(s)

NORTHROP GRUMMAN

- What is the danger of dropping A variable?
  - Was there a logical/theoretical reason for including the variable? Have other programs, projects or studies exhibited A relationship?
    - If the collinear variables will always move together then dropping one of them or replacing them with a more comprehensive variable may not create any problems
    - If they will in the future move in opposite directions then dropping one of them would be A misspecification of the relationship which may result in forecasts of the dependent variable that are far from the mark

Gather More Data

NORTHROP GRUMMAN

- Most of the time we are already using all of the data that is available
  - If there is more data available determine the impact of adding the additional data

Gather More Data

NORTHROP GRUMMAN

- Results with ten data points

Notional Aircraft Program		
	QTY	*MIDPOINT*
QTY	1.00000	
MIDPOINT	0.78886	1.00000

Gather More Data

NORTHROP GRUMMAN

- Notional Aircraft Program Lot Structure

CONTRACT	QTY	MIDPOINT
FSD	14	5.6
PILOT	9	18.8
LIMITED	30	37.3
FY81	74	87.3
FY82	88	169.1
FY83	126	275.5
FY84	135	406.7
FY85	146	547.5
FY86	139	690.6
FY87	109	815.3
FY88	93	916.5
FY89	84	1005.1
FY90	84	1089.2
FY91	70	1166.3
FY92	48	1225.4

Gather More Data

NORTHROP GRUMMAN

- Results With Fifteen Data Points

Notional Aircraft Program		
	QTY	*MIDPOINT*
QTY	1.00000	
MIDPOINT	0.26900	1.00000

Combine The Collinear Variables Into A New Variable

NORTHROP GRUMMAN

- Replace Y = a1 + b1*X1 + b2*X2 With

$$Y = a2 + b3*X3$$

Where:

$$X3 = k1*X1 + k2*X2$$

The values of k1 and k2 represent the Relative relationship between Y, X1 and X2

What Are Some Non-standard Solutions?

NORTHROP GRUMMAN

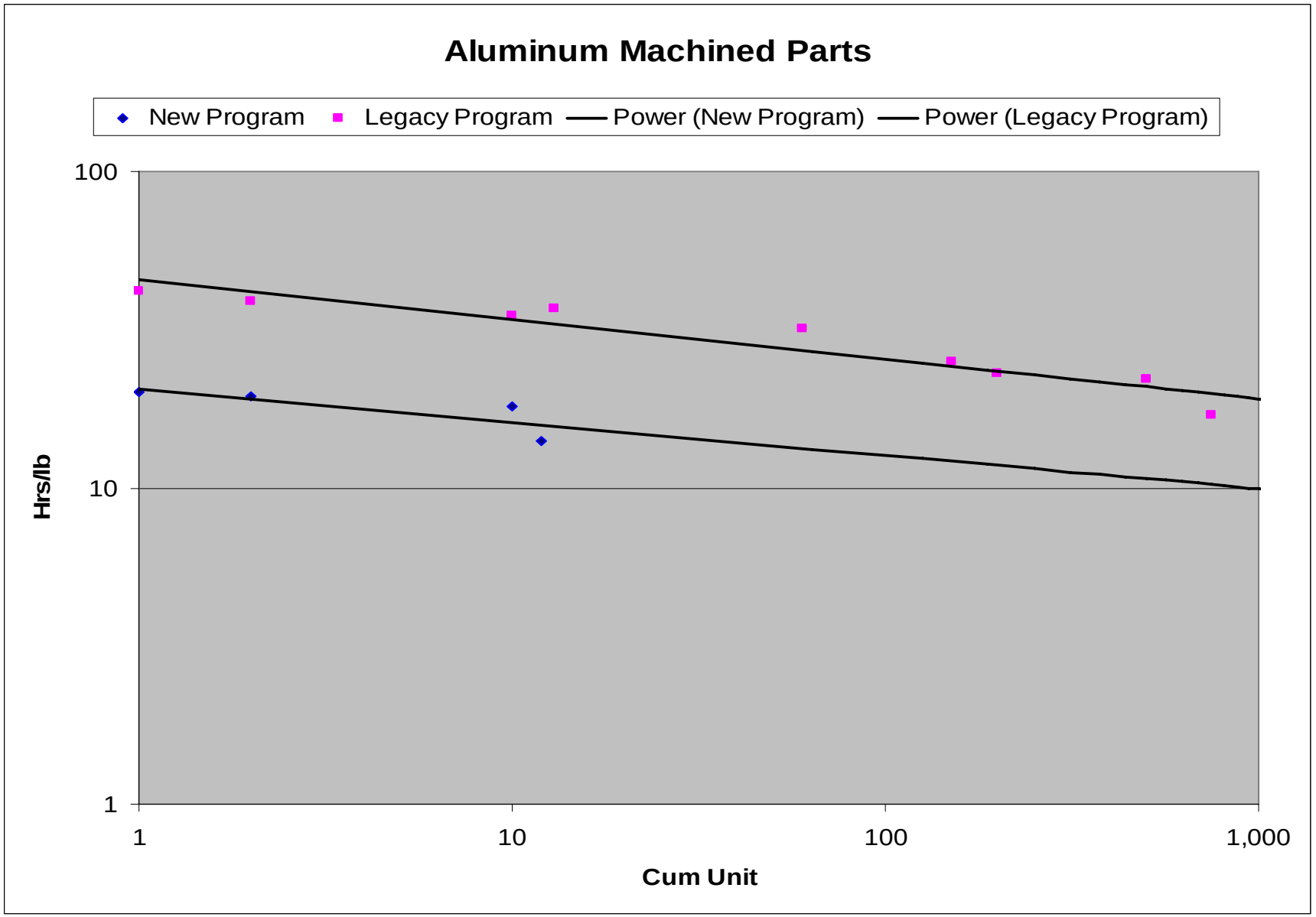
- Gather More Data Sort Of
- Normalize The Data
- Ridge Regression

Gather More Data

NORTHROP GRUMMAN

- If additional data is not available is analogous data available?
  - Are there similar programs/projects that may share the same relationships or relative relationships?
  - Consider pooled regression analysis where some of the relationships are common between programs/projects within processes
    - Costs of A common process for two programs/projects may share an improvement curve slope and A weight advantage curve (aka ARCO curve), but have different t-1s

Gather More Data



Normalize The Data

NORTHROP GRUMMAN

- Consider adjusting the data for the “offending” variable using A known relationship from other studies

$$Yadj = Y \text{ (actual values)} - b1*X1 \text{ (actual values)}$$

Where: b1 is a coefficient borrowed from another program, project or accepted industry value. An example of this might be an improvement curve slope or A weight advantage curve.

- Then estimate the equation:

$$Yadj = a1 + b2*X2$$

Ridge Regression Analysis

NORTHROP GRUMMAN

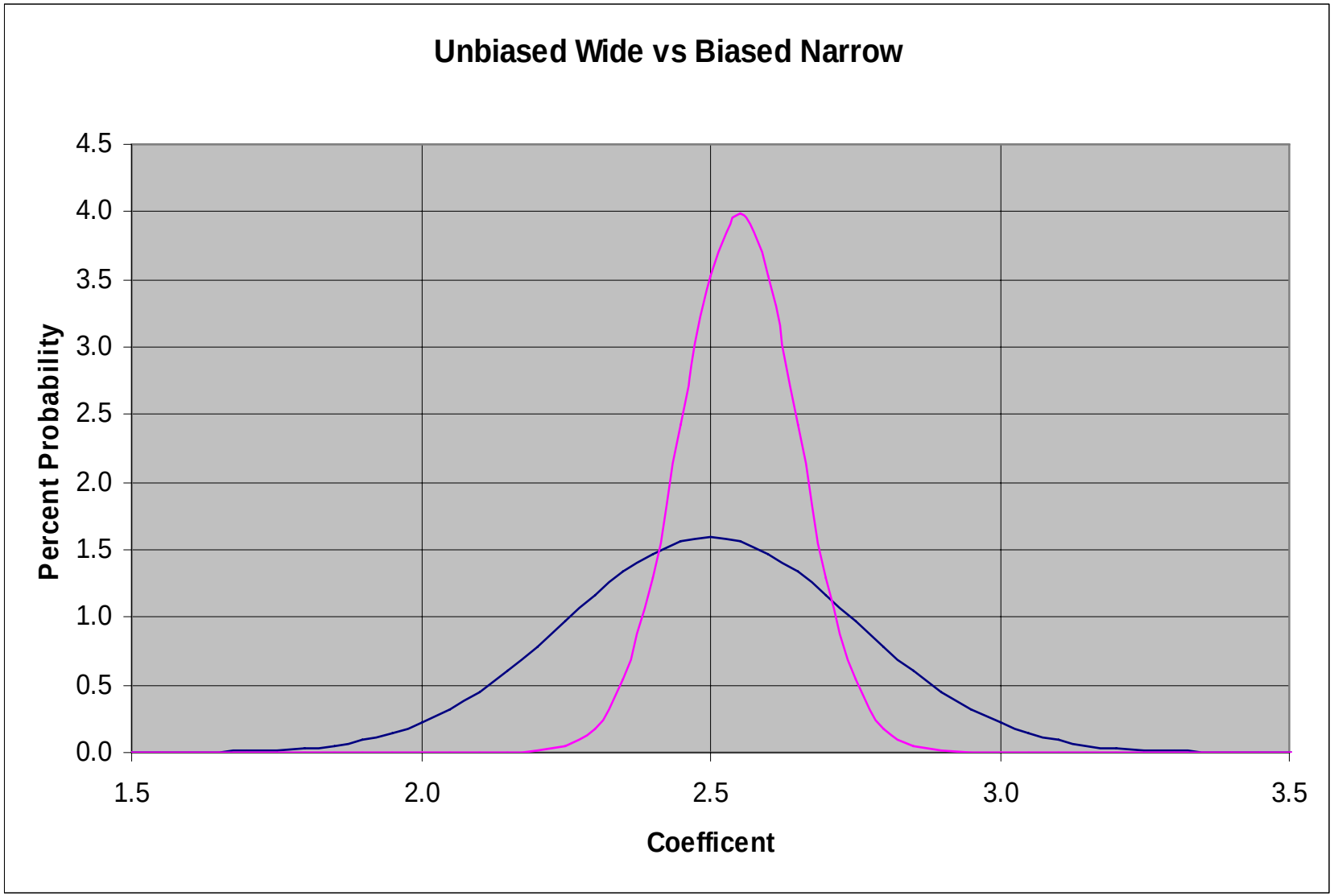
- What is ridge regression?
  - It is an approach for solving the problem of multicollinearity that introduces bias into the estimating process in an effort to achieve greater efficiency, i.e. Less dispersion in the error terms of the regression
  - It achieves this by raising the major diagonal in the $[X'X]^{-1}$ matrix, this is where the ridge comes from

Note: in Ordinary Least Squares Regression Analysis the vector of coefficients, b, is:

$$b = [X'X]^{-1}X'Y$$

Ridge Regression Analysis

NORTHROP GRUMMAN



Summary - Options

NORTHROP GRUMMAN

- Drop The Offending Variable(s) Or Replace It (Them) With A More Definitive Variable(s)
- Gather More Data
- Combine The Collinear Variables Somehow Into A New Variable
- Gather More Data Sort Of
- Normalize The Data
- Ridge Regression

Summary

NORTHROP GRUMMAN

- Before dropping an independent variable because its coefficient isn't statistically significant or has the wrong sign ask the question:
  - Did I include this variable because it makes sense from A logic and A theoretical standpoint? Have other programs, projects or studies exhibited A relationship?
    - If the answer is no - then it probably shouldn't be a part of your analysis anyway
    - If the answer is yes - then explore alternatives but include the usual statistical testing to be sure you aren't seeing something that really isn't there

NORTHROP GRUMMAN

DEFINING THE FUTURE