



Automated Cost Estimating Integrated Tools

# Getting a Good Grip on Grouping

2008 ACEIT User Conference

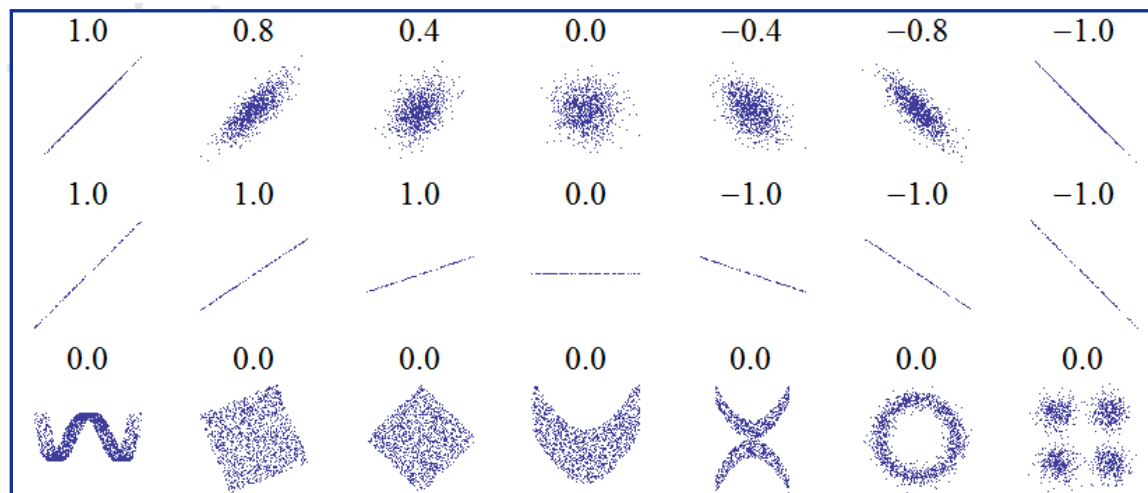




- **What is RI\$K grouping?**
- **4 of the 5 Steps for Grouping**
- **The Grouping and Correlation Wizard**
- **Step 5: Choosing Strengths**
- **ACE Grouping is NOT Spearman Rank Correlation!**
- **Questions**

# What is Correlation?

- ACE RI\$K grouping is an informal way of correlating rows.
- But what is correlation?
  - Basically, “correlation” is a pairwise relationship where the values of two elements would likely increase or decrease together.
    - An “inverse correlation” is where the values may move in opposite directions.
  - We often use statistical correlation to find cost predictors (i.e. a CER).
  - The two most popular correlation measures in statistics are Pearson Product Moment and Spearman Rank.



*Image from <http://en.wikipedia.org/wiki/Correlation>*



# “Relationship” Heuristic

## ■ ACE Help Text:

- *“In keeping with the ‘simple is better’ principle, we capture the impact of correlation between risk elements using a heuristic approach.”*

## ■ “RI\$K group strength” is a heuristic for defining relationships among a number of line items in the estimate.

- The term “heuristic” is analogous to a cognitive inference.
  - It finds—or at least gets near—the answer quickly.
- We use “strength” instead of “correlation” to alleviate confusion, as “correlation” has a specific connotation in statistics.

## ■ ACE doesn’t deal with the complications of a full “input” correlation matrix.

- Often, existing algebraic relationships, such as shared input drivers, already capture correlation and no “input” correlation is required—***or even desired.***
- Grouping adjusts the relationships among the uncertainty of a collection of rows in the session to produce the desired “output” correlation.



# PMCC Range Inconsistency

- The example below shows the valid coefficient ranges for PMCC\*.
  - Ranges change depending on the shapes of the two distributions.
  - Note: the shape of a Log Normal distribution changes with its dispersion.

Distribution Form	Normal (CV*=.2)	Uniform (CV=.2)	Log Normal (ASE*=.1)	Log Normal (ASE=.5)
Normal (CV=.2)	+/-1.0	<b>+/- 0.978</b>	+/- 0.997	+/- 0.938
Uniform (CV=.2)		+/-1.0	+/- 0.974	+/- 0.898
Log Normal (ASE=.1)			1.0 to -0.988	<b>0.960 to -0.912</b>
Log Normal (ASE=.5)				<b>1.0 to -0.778</b>

*Values are approximate*

- A PMCC of 0.9 is **invalid** for some mixtures of distributions.
  - This produces an inconsistent “Input” correlation matrix.
  - A correlation report generated from a calculation result could never show 0.9 for those mixtures—*which may go counter to documented recommendations.*

*\* PMCC = Pearson Product Moment Correlation Coefficient, CV = Coefficient of Variation, and ASE = Adjusted Standard Error*



# Keeping it Consistent

- **Conversely, Group Strength is always defined from -1.0 to 1.0.**
  - In that way, even if further analysis of a row requires a change in uncertainty (e.g., altering the ASE), the assigned “strength” of the relationship isn’t lost.
  - ACE will automatically adjust the strength to fall within a valid PMCC range when correlating Latin Hypercube input draws.
- **Any combination of strengths for a group is always valid.**
  - This is due to the fact that ACE uses a coefficient vector instead of a full matrix.
- **However, certain combinations coefficients may make an “Input” PMCC matrix *inconsistent*.**
  - Is this correlation matrix consistent?
  - Are any coefficients out of range?

<i>Contrived “Input” PMCC Matrix</i>	Row A	Row B	Row C	Row D
Row A	1	0.95	-0.75	0.85
Row B		1	0.75	0.20
Row C			1	1.00
Row D				1



# Gauging Relationship Strength

- **Table 3-1 offers guidance on interpreting coefficients generated from results.**
  - We're interested in the results because those coefficients relate to regression data—NOT an “input” matrix.
- **You can use this table when applying group strengths (in step 5).**
  - Use the values in table 3-1 if you have a dominant row defined in your group.
  - Without a dominant row, you would take the square root of the factor of interest.
- **The PMCC for two rows is equal to...**
  - ...the product of the group strengths assigned to the two rows...
  - ... but only if both rows have Normal distributions, do not share inputs, and all inputs used on the rows are not correlated.

**Table 3-1** Default Correlation Factors

Strength	Positive	Negative
None	0.00	0.00
Weak	0.25	-0.25
Medium	0.50	-0.50
Strong	0.90	-0.90
Perfect	1.00	-1.00

*From The 2007 AFCAA Cost Risk and Uncertainty Handbook*

Two Rows With ACE Group Strengths of ...	Theoretical PMCC
0.9 and 1.0	0.9 (Strong)
0.707 and 0.707	0.5 (Medium)
-0.5 and 0.5	-0.25 (Weak)
0.95 and 0.95	0.9 (Strong)

# 4 of 5 Steps For Grouping





# Applying Grouping

- *(Step -3: Build your model – point estimate)*
- *(Step -2: Apply uncertainty to rows)*
- *(Step -1: Resist applying grouping before you examine results.)*
  
- **Step 1: Make sure you need it.**
- **Step 2: Find discrepancies.**
- **Step 3: Fix the errors.**
- **Step 4: Choose where to apply it.**
- **Step 5: Apply group strengths.**



## Step 1: Make Sure You Need It

- **Parametric models have correlation integrated in them.**
  - From shared technical parameters (e.g. Learning Slope or Overhead)
  - From shared input drivers (e.g. two CERs using Weight)
  - From algebraic build-ups (e.g.  $0.2 * PMP\$$ )
  - From roll-ups (e.g. parents are correlated to their children)
- **Before you apply RI\$K grouping, look for existing correlation present in model. [1]**
  - Run a RI\$K Correlation report on rows of concern.
    - You can filter by a category column to organize contents and narrow matrix size.
    - Note: The RI\$K Correlation report reports PMCC and not group strength.
  - Find rows where you feel the coefficient is too high or too low.
    - In my opinion, if the coefficient is within +/- 0.1 of expected value, leave it alone.
  - Verify that no unintended consequence is causing the discrepancy.
    - Is an input driver being shared that shouldn't be? (e.g. a convenient variable)
    - Should an input driver be shared that isn't? (e.g. two different weight measures)
    - Is risk being applied to a row that shouldn't have it? (e.g. a case input)

[1] *The AFCAA Cost Risk and Handbook offers detailed procedures on finding and correcting errors*



## Step 2: Find Discrepancies

- **Look for relationships that you would expect to see.**
  - For instance...
    - It makes sense that PMP\$ and SEPM\$ should be highly correlated.
    - Production and maintenance costs for an engine should likely be correlated.
    - Engine thrust and payload weight should usually be correlated.
- **Look for relationships that you wouldn't expect to see.**
  - For instance...
    - It makes sense that labor costs for research would not be strongly correlated to labor costs for operations if performed by different companies or different staff. Moderate correlation may make sense.
    - Production site costs are likely uncorrelated to the cost of the navigation system.
    - Uncertainty for labor rates, learning slopes, etc. across different commodities (different companies) are likely not strongly correlated.



## Step 3: Fix The Errors

- **Are you sharing a variable out of convenience?**
  - For instance... Early on, you may have defined a single variable for Learning Slope and used it on all your learning rows even unrelated ones.
- **Are technical inputs uncorrelated that you expect to be?**
  - For instance... You might have defined the same weight in different units on two different rows.
  - Or, maybe Slope should be related to a “complexity” factor used elsewhere.
- **Usually, uncertainty is not applied to “what-if” case inputs.**
  - Do not put a huge dispersion on an input that is defined in “what-if” case (e.g., user may enter of buy quantity of 10, 20, or 30 for different drills – that doesn’t mean that the “uncertainty” for BQ ranges from 10 to 30!)
  - Apply uncertainty that relates to risk that exists after the decision was made.
    - For instance, you know that payload may increase due to requirements creep.
    - Or, technology needed for O&S may become cheaper several years from now.
- **A shared input without uncertainty can’t correlate rows using it.**
  - This may be symptomatic of a missing uncertainty on that input row.



## Step 4: Choose Where to Apply It

- **A 0.7 Group Strength** (*unfortunately*) **doesn't add 0.7 to a coefficient.**
  - A get the most impact, find the rows whose uncertainty is dominant.
- **Direct your attention to unshared input drivers among costs.**
  - Adding correlation to input drivers will likely permeate through model in ways that correlation on specific costs rows will not.
  - Uncertainty on inputs compound uncertainty at later stages of the model's calculation – giving more bang for the buck.<sup>[2]</sup>
- **Recognize relative uncertainty of a CER and its inputs.**
  - When a CER has a large uncertainty relative to its inputs...
    - The correlation between this CER and other rows will not increase much when inputs are grouped.
    - In this case, you will likely have to group the CER instead of its inputs.
  - If the CER is close to certain compared to its input's uncertainty...
    - The correlation of the CER will increase significantly when inputs are grouped.
    - In this case, grouping the CERs will have little impact on their correlations.

*[2] Smith (2007) Functional Correlation paper shows the compounding effect compared to throughputs*



# Validate Routinely

- **After making changes, analyze the new results.**
- **Your changes will likely propagate through the model and “repair” other discrepancies for you automatically.**
- **Routine fix-and-validate saves you from accidentally overcompensating.**



# Basic Risk Example

- In the “02a – Basic Risk” example session, the two technical characteristics are grouped together.
  - When the *Weight* goes up, the *Range* goes down.
  - Therefore, a negative group strength is appropriate.
  - In this case, the relationship is considered “moderate” in recognition that the uncertainty of *Range* is impacted by more than simply the *Weight*.
  - Note that *Weight* is “dominant,” signified with *D*—i.e., it drives *Range*.

ACE 7.1 - [02a - Basic Risk.aceit (Read-Only) - RISK Basic (BY2006\$K)]

File Edit View Documentation Calc Cases Reports Tools Window Help

Arial 10 B I U \$00 RISK Basic

33 -0.8

	WBS/CES Description	Unique ID	Point Estimate	Equation /	Distribution Form	Grouping	Group Strength
30							
31	*Technical/Performance Characteristics						
32	Air Vehicle Takeoff Weight (lbs)	TW	12000.00 (25%) *	12000	Triangular	PERF	D
33	Air Vehicle Range (nmi)	RANGE	250.00 (42%) *	250	Triangular	PERF	-0.8
34							
35							

Ready NUM



# Grouping and Correlation Wizard





# Grouping Wizard

- ACE has a dialog to help group rows. (found on the Tools menu)
- You can create, edit, switch and view groups.
- Strengths are defined as a vector of coefficients.
- ACE displays a theoretical PMCC matrix for reference.
  - Matrix values assume all distribution forms are Normal with uncorrelated inputs

**RISK Grouping and Correlation**

Selected Grouping

Group ID:

1. Assign "D" to the dominant element in the Group.  
2. Assign a value of 0.0 (none) to 1.0 (complete) correlation to each Group member.  
3. Otherwise, use "Assign Correlation of" utility to create a matrix that contains the same value.

	WBS/CES Description	Total	Strength	64	65	150	200	201	205
64	Navigation/Guidance	2.647 (42%)*	0.7071	1.000	0.500	0.387	0.500	0.500	0.500
65	Propulsion	5.723 (42%)*	0.7071		1.000	0.387	0.500	0.500	0.500
150	Basic Structure T1	2.194 (50%)*	0.5477			1.000	0.387	0.387	0.387
200	Propulsion Unit Cost	5.127 (50%)*	0.7071				1.000	0.500	0.500
201	Transportation Unit Cost	0.599 (17%)*	0.7071					1.000	0.500
205	Block Buy Cost at	7.717 (15%)*	0.7071						1.000



# Step 5: Choosing Strengths





# Finding a Dominant Element

- **Matrix is from the “Costat Helicopter” example pairwise report.**
  - Examine pairwise correlations between technical parameters being used.
- **Suppose our model contains CERs from this data set that use...**
  - Shaft Horsepower (SHP), Weight (WGHT), Range, Climb Rate (ClimbRt), Empty Weight (TW1)
- **Dominant prospects have high coefficients:**
  - *WGHT (0.87)*, *SHP (0.85)*, *TW1 (0.86)* are potential dominant items.
  - Maybe use *WGHT* because it seems to “drive” the other elements.

*Correlation matrix from CO\$TAT report*

Pairwise Variable Analysis For Dataset Heli UC									
	RATE	LEN	WGHT	SPEED	RANGE	ClimbRt	SHP	TW1	AF_UC
RATE	1.0000	-0.1312	-0.4910	-0.1099	-0.4897	-0.0015	-0.5225	-0.5171	-0.6376
LEN	-0.1312	1.0000	0.9728	0.3842	0.8131	0.6336	0.9860	0.9634	0.7748
WGHT	-0.4910	0.9728	1.0000	0.2600	0.8417	0.7009	0.9819	0.9711	0.7931
SPEED	-0.1099	0.3842	0.2600	1.0000	0.2355	-0.0047	0.4230	0.4086	0.3735
RANGE	-0.4897	0.8131	0.8417	0.2355	1.0000	0.3263	0.8229	0.8996	0.8890
ClimbRt	-0.0015	0.6336	0.7009	-0.0047	0.3263	1.0000	0.6313	0.5753	0.3982
SHP	-0.5225	0.9860	0.9819	0.4230	0.8229	0.6313	1.0000	0.9839	0.8247
TW1	-0.5171	0.9634	0.9711	0.4086	0.8996	0.5753	0.9839	1.0000	1.0000
AF_UC	-0.6376	0.7748	0.7931	0.3735	0.8890	0.3982	0.8247	0.8701	1.0000
<b>Average of coefficients of interest:</b>			<b>0.8739</b>		<b>0.7226</b>	<b>0.5584</b>	<b>0.8550</b>	<b>0.8575</b>	

Copy this vector into Group Strength col. Assign “D” to WGHT.



# Result From Paste

- It is easier to define a “dominant row” for the group using “D” for the cost driver for the group.
  - A dominant row lets you copy directly from a PMCC matrix into the Strength column without taking the square root of coefficients.
  - Without a dominant row, you need to do some math.
  - “D” has an assumed Group Strength of 1.0.
  - *(You could also use 1.0 for WGHT with the same result – except that you will no longer be able to define a random seed for the group. ACE will use the “D” row’s seed for group or else create an internal one if no “D” exists.)*

	WBS/CES	Total	Strength	133	134	135	136	137
133	WGHT		D	1.000	0.842	0.701	0.982	0.971
134	RANGE		0.8417		1.000	0.590	0.826	0.817
135	ClimbRt		0.7009			1.000	0.688	0.681
136	SHP		0.9819				1.000	0.954
137	TW1		0.9711					1.000



# Ripple Effect of Strengths

- You can calculate the theoretical PMCC = product of the two strengths.
  - Note: ACE fills matrix values by assuming Normal distributions and uncorrelated inputs.
  - The actual PMCC for the rows will **decrease** if the distributions are not Normal.
  - The actual PMCC for the rows will **increase** if they share input drivers.
- For example,
  - RANGE has a strength of **0.8417** and SHP has a strength of **0.9819**, giving a theoretical PMCC of **0.826**.
  - The non-dominant coefficients will always be less than or equal to the group strengths.

	WBS/CES	Total	Strength	133	134	135	136	137
133	WGHT		1.000	1.000	0.842	0.701	0.982	0.971
134	RANGE		0.8417		1.000	0.688	0.826	0.817
135	ClimbRt		0.7009			1.000	0.688	0.681
136	SHP		0.9819				1.000	0.954
137	TW1		0.9711					1.000



# Result from Applying Strength

- **Green is Group Strength.**
  - ACE uses only one vector from correlation matrix.
- **Orange is “off-axis” correlation.**
  - Second chart shows secondary effect on correlation from group strength.
  - *For simplicity, All variables have Normal uncertainty.*
- **Third chart shows difference between data and ACE.**
  - Note only the “off-axis” elements miss their target.
  - Only [ClimbRt vs. Range] has notable difference in correlation.
- **The differences will likely be lost in the “noise” of the model.**
  - No further adjustment is needed

Original Correlation Matrix from CO\$TAT					
	WGHT	RANGE	ClimbRt	SHP	TW1
WGHT	1.000	0.842	0.701	0.982	0.971
RANGE		1.000	<b>0.326</b>	<b>0.823</b>	<b>0.900</b>
ClimbRt			1.000	<b>0.631</b>	<b>0.575</b>
SHP				1.000	<b>0.984</b>
TW1					1.000

ACE Theoretical					
	WGHT	RANGE	ClimbRt	SHP	TW1
WGHT	1.000	0.842	0.701	0.982	0.971
RANGE		1.000	<b>0.590</b>	<b>0.826</b>	<b>0.817</b>
ClimbRt			1.000	<b>0.688</b>	<b>0.681</b>
SHP				1.000	<b>0.954</b>
TW1					1.000

Difference between Data and ACE (Theoretical) PMCC					
	WGHT	RANGE	ClimbRt	SHP	TW1
WGHT	0%	0.0%	0.0%	0.0%	0.0%
RANGE		0%	<b>26.4%</b>	0.3%	-8.3%
ClimbRt			0%	5.7%	10.6%
SHP				0%	-3.0%
TW1					0%



# ACE Grouping Is NOT Spearman Rank Correlation





# NOT Spearman Rank!!!

- **Some people mistake ACE Grouping with Spearman Rank Correlation...**
  - ...since the coefficient range is always -1.0 to 1.0.
  - ...and they saw the word “ranking” mentioned in the algorithm.
  
- **This conclusion is WRONG.**
  
- **ACE uses non-linear version-inversion transforms to convert from PMCC using assumption of normality to the appropriate inverse CDF while applying constraints of Latin Hypercube draw selection, penalty factors, and truncation.**
  - The algorithm is summarized in ACE Help.<sup>[3]</sup>

*[3] Smith-Hu (2003) white paper details the derivation from the Lurie-Goldberg correlation algorithm*



# Conclusion

- Correlation measures the relationship between the movements up and down of two elements' values.
- ACE RI\$K Group Strength is a heuristic used for adjusting the resulting correlation of rows—based on Pearson.
- Use the *RI\$K Correlation* report (calculated results) when evaluating correlations coefficients.
- Only group rows when correlation is not already captured by the functional relationships inherent in the model.
- Watch for errors that produce unwanted correlation.
- The *Grouping and Correlation Wizard* simplifies creating and maintaining RI\$K groups.
- Strength is neither Spearman Rank nor strictly Pearson Product Moment.



THANK YOU

