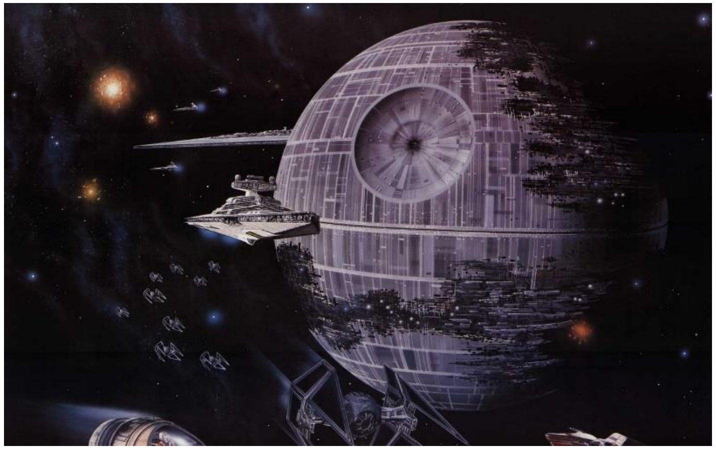
A Significant Other: Hypothesis Testing on Qualitative Predictor Variables in CER Development

Cassandra Chang, Program Planning & Control Office
Glenn Research Center

May 4, 2023

Outline

- Background on CERs
- Types of variables and variable selection
  - Qualitative vs quantitative variables
- Hypothesis testing on regression coefficients
  - Simple linear regression vs nonlinear regression
  - Hypothesis formation and decision rules
- Working the hypothesis tests in practice
  - Example with CO$TAT output
- Conclusion and Questions

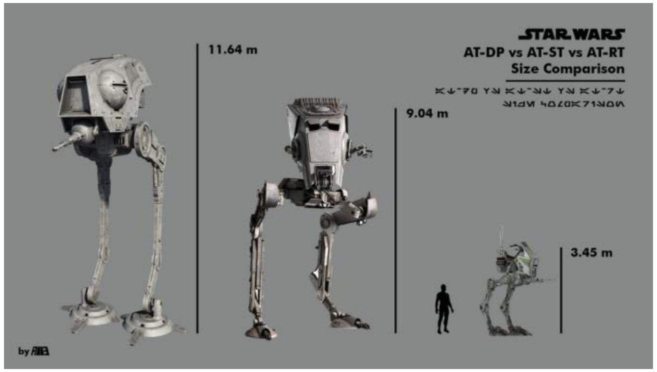


Why CERs?

- CER = Cost Estimating Relationship
- Cost estimates based on historical data
- Understand the equations
  - Data used to train the model
  - Assumptions applied
- Other cost estimating tools use their own CERs but we have no insight on how those were developed
  - Which are the driving variables?
  - Were outliers excluded and why?

Variable Selection

- Available data: technical data from CADRe, cost normalization data from RedStar
  - Both quantitative and qualitative data available
  - We will be focusing on *qualitative* data



- Quantitative data: continuous
  - Mass, weight, height

- Qualitative data: discrete categories
  - Material, weight class, color

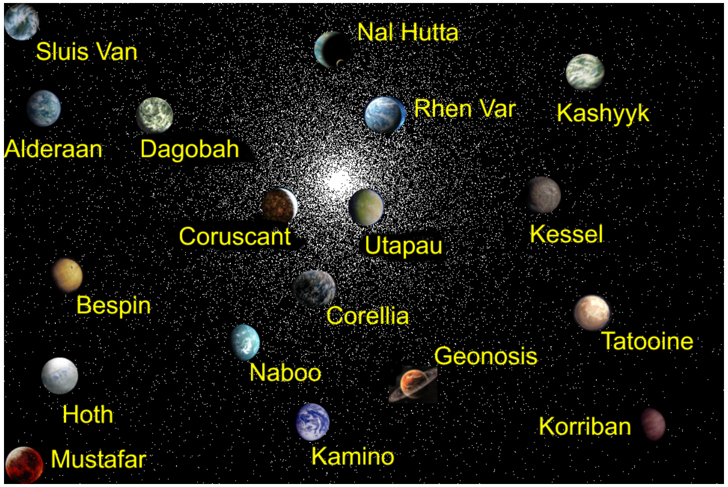
Variable Selection

- Use variables that make sense
  - Variables can be statistically significant but not practically significant
    - What is causing this variable to be statistically significant?
    - What variables logically should affect total cost?
  - What parameters will be we able to input when applying to new predictions?

- Treat variables in a way that makes sense
  - Should still be able to interpret the regression results
  - What do transformations mean?

Qualitative Variables

- Many of the desired parameters for previous space missions are not quantitative
  - Earth orbiting vs planetary
  - Mission class
  - Type of power system



- Best coded in a binary system
  - 0 for absence of trait, 1 for presence of trait
  - For multiple levels, use a different variable for each level to avoid assumptions about level effects
    - Using numerical labels will force the assumption that each category is spaced apart equally when that may not be the case
    - The change from Class A to Class B is not necessarily the same as the change from Class B to Class C

Equation Forms

Let X_1 be a quantitative variable and X_2 be a qualitative binary variable.

Simple Linear Regression:

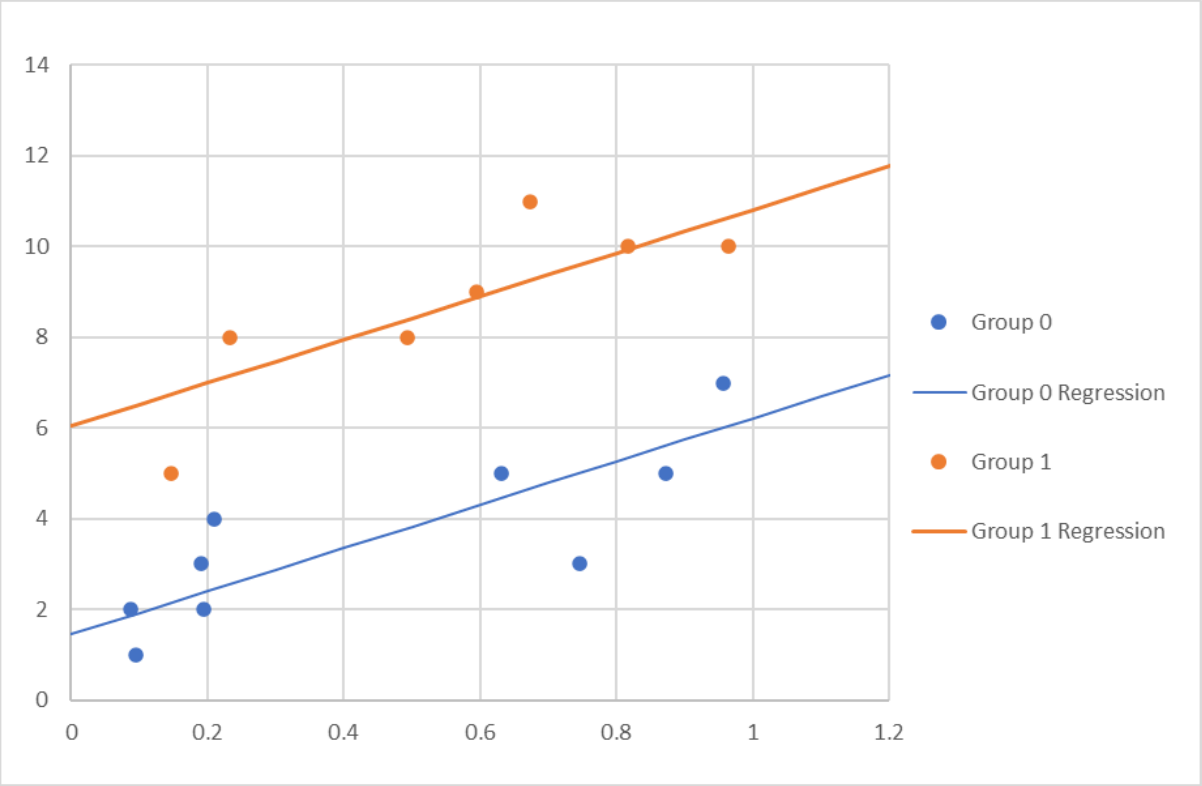
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Nonlinear Regression:

$$Y = \beta_0 {x_1}^{\beta_1} {\beta_2}^{x_2}$$

Linear Regression Equation Form

Simple Linear Regression Example: $Y = 1.44 + 4.77X_1 + 4.59X_2$



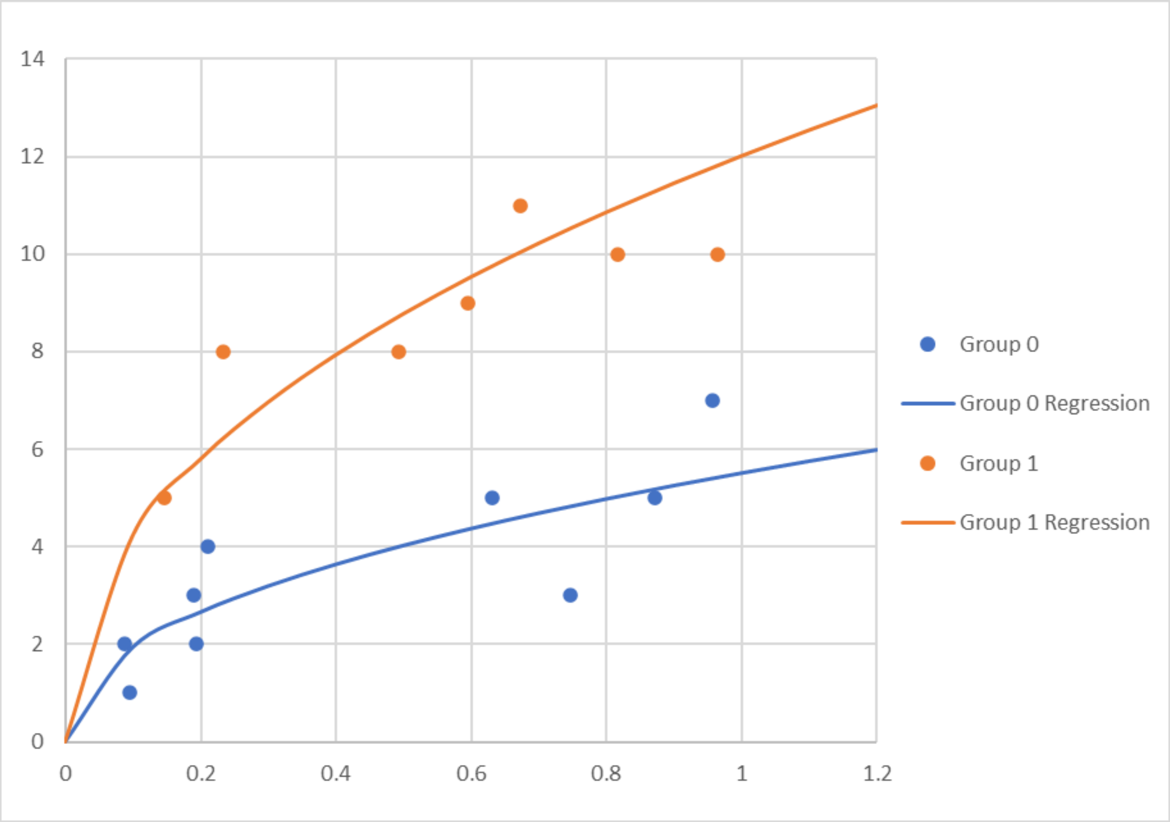
When $X_2 = 0$:
$Y = 1.44 + 4.77X_1$

When $X_2 = 1$:
$Y = 6.03 + 4.77X_1$

Presence of X_2 is an additive term

Nonlinear Regression Equation Form

Nonlinear Regression Example: $Y = 5.52X_1^{0.45}\ 2.17^{X2}$



When $X_2 = 0$:
$Y = 5.52X_1^{0.45}$

When $X_2 = 1$:
$Y = 12.01X_1^{0.45}$

Presence of X_2 is an multiplicative factor

Variable Significance for Linear Regression

In general, what we have traditionally been testing our variables at are:

- Null hypothesis: $\beta = 0$
  - This works for simple linear regression
- Alternate hypothesis: $\beta \neq 0$

- Two tailed test: could be either greater than or less than zero, even if one of those options does not make practical sense.
- In simple linear regression, this works for both continuous and binary variables. This is not the case for nonlinear regression!

Variable Significance for Nonlinear Regression

Instead, we need to take a look at the coefficients as factors that affect the cost.

- Null hypothesis: $\beta = 1$
  - A factor of 1 shows no correlation
- Alternate hypotheses:
  - If the factor is expected to have negative correlation with the total cost, $\beta < 1$
  - If the factor is expected to have positive correlation with the total cost, $\beta > 1$
- One tailed tests – use subject matter expertise to determine expected impacts
  - For example, planetary missions should cost more than earth orbiting missions

A Practical Application

Model Form:	Weighted Non-Linear Model
Non-Linear Equation:	Y = 0.621 * X1 ^ 0.8962 * 1.786 ^ X3 * 2.206 ^ X4
Error Term:	MUPE (Minimum-Unbiased-Percentage Error)
Minimization Method:	Gauss-Newton

Consider a data set with 43 observations.

- Variables considered:
  - Y = total cost in FY14$M
  - X1 = mass in kg
  - X3 = 1 if planetary, 0 if earth orbiting
  - X4 = 1 if flagship mission, 0 if not
- Regression performed using MUPE nonlinear regression in CO$TAT (part of the ACEIT suite - thanks Tecolote!)
- Is the mission type (planetary or earth orbiting) a statistically significant predictor of cost?

A Practical Application

Model Form:	Weighted Non-Linear Model
Non-Linear Equation:	Y = 0.621 * X1 ^ 0.8962 * 1.786 ^ X3 * 2.206 ^ X4
Error Term:	MUPE (Minimum-Unbiased-Percentage Error)
Minimization Method:	Gauss-Newton

Let β_3 be the coefficient associated with X3.

- Hypotheses:
  - Null hypothesis: $\beta_3 \leq 1$
  - Alternate hypothesis: $\beta_3 > 1$
- Decision rule:
  - If $t^* \leq t(1-\alpha; n-5)$, do not reject the null hypothesis
  - If $t^* > t(1-\alpha; n-5)$, reject the null hypothesis
    - (p-value < 0.05)

A Practical Application

Variable/Term	Coefficient Estimate	Approximate Std Error	Approximate Lower 95% Confidence	Approximate Upper 95% Confidence	P-Value	Test Type
a	0.6210	0.1932	0.2303	1.0117	0.001310717	one tail > 0
b	0.8962	0.0835	0.7274	1.0650	1.6368E-13	one tail > 0
c	1.7860	0.2651	1.2499	2.3221	0.002567809	one tail >1
d	2.2057	0.4466	1.3023	3.1091	0.005106907	one tail >1

- Test statistic calculated by $t^* = \frac{c-1}{s\{c\}}$ where c is the binary variable
- P-value in Excel calculated using the formula:
  - TDIST(CoefficientEstimate/StdError, DegreesofFreedom, NumberofTails)
  - In this case, p-value for c uses the formula

    =TDIST((1.786-1)/0.2652, 39, 1)

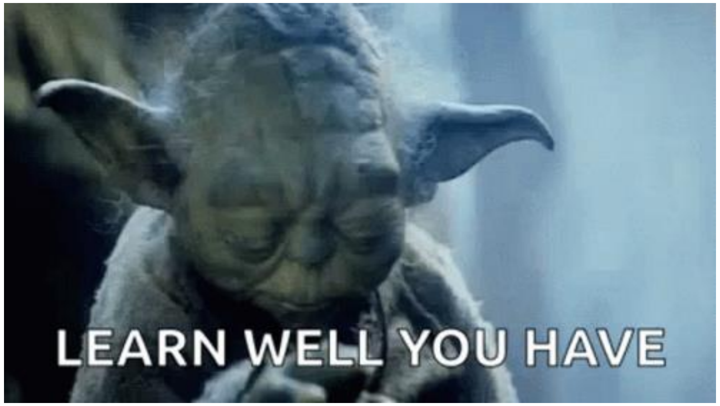
Some Food for Thought

A few other notes:

- Specify significance level
  - Although a 95% confidence level, or $\alpha = 0.05$, is a typical standard, it is not the only standard that can be applied
- Note any outliers
  - Outliers or missions can be left out if they appear to be less relevant
  - Keep track of why missions were excluded or what subgroups the data set was split into
- Clearly state input parameter ranges
  - What are the minimum and maximum parameter inputs used
  - Additional uncertainty may be needed if the input parameter is outside the range of the data set

Conclusion

- CERs are powerful tools when we understand them and apply them appropriately
- It is important to note the assumptions made when performing regression and other prediction techniques
- Qualitative variables need to be treated differently than quantitative variables when hypothesis testing
- Use logic and subject matter expertise!
- Learn from my mistakes ☺



Thanks for listening!

Questions?