Regression and interpretation low R-squared!

Social Research Network 3nd Meeting

Noosa

April 12-13, 2012

Kenshi Itaoka

Mizuho Information & Research Institute, Inc.

Contents

- Motivation
- About r
- Purpose of regression
- Example
- Conclusion

Motivation

- We sometime encounter low R squared (Fitness function in regression) in our results of a regression analysis.
- We sometime hear it is not possible or not appropriate to draw any meaningful insights from regression analyses because the R squared is very low.
- Is this true?

Fitness function in regression

- R-squared= (1- SSE) / SST

  Defined as the ratio of the sum of squares explained by a regression model and the "total" sum of squares around the mean.

  Interpreted as the ration of variance explained by a regression model

- Adjuseted R-squared= (1- MSE) / MST

  MST = SST/(n-1) MSE = SSE/(n-p-1)

- Other indicators such as AIC, BIC etc. also sometime used for model selection.

Purpose of regression

- Analysis of partial correlation between factors: to avoid risks of interpretation of simple correlations in multi-variable analyses
  - Examination of influential factors on phenomena to be explained
- Causal inference? Yes / No: need to check of logic of causality.

Purpose of regression 2

- Forecast
  - Example: temperature and energy consumption
  - Predictors (independent variables)
  - Prediction (dependent variables)

- Detection of influence of internal /external factors Internal consistency check
  - Example: demographics and public opinions
  - Explanatory variables (independent variables)
  - Explained variable (dependent variables)

Model estimation in regression

- True model*: y = b0 + b1 x1 + b2 x2 + u*
- Estimated model: *y = a0 + a1 x1 + v*
  - X1 and x2 should be independent
  - No correlation between x1 and v as the theory of the least square methodology

In this case,

a1 is equal to *b1*

But variance of error term is influenced by *x2*

Variance of *u* is smaller in than in *v* by *x2*

R squared should be smaller in the estimated model than that in the true model due to *x2.*

- To examine the effectiveness of a1, the size of R squared does not matter.
- Only significance of at matter
- In practice, examination of correlation between *x1* and *x2* (if observed) is important.

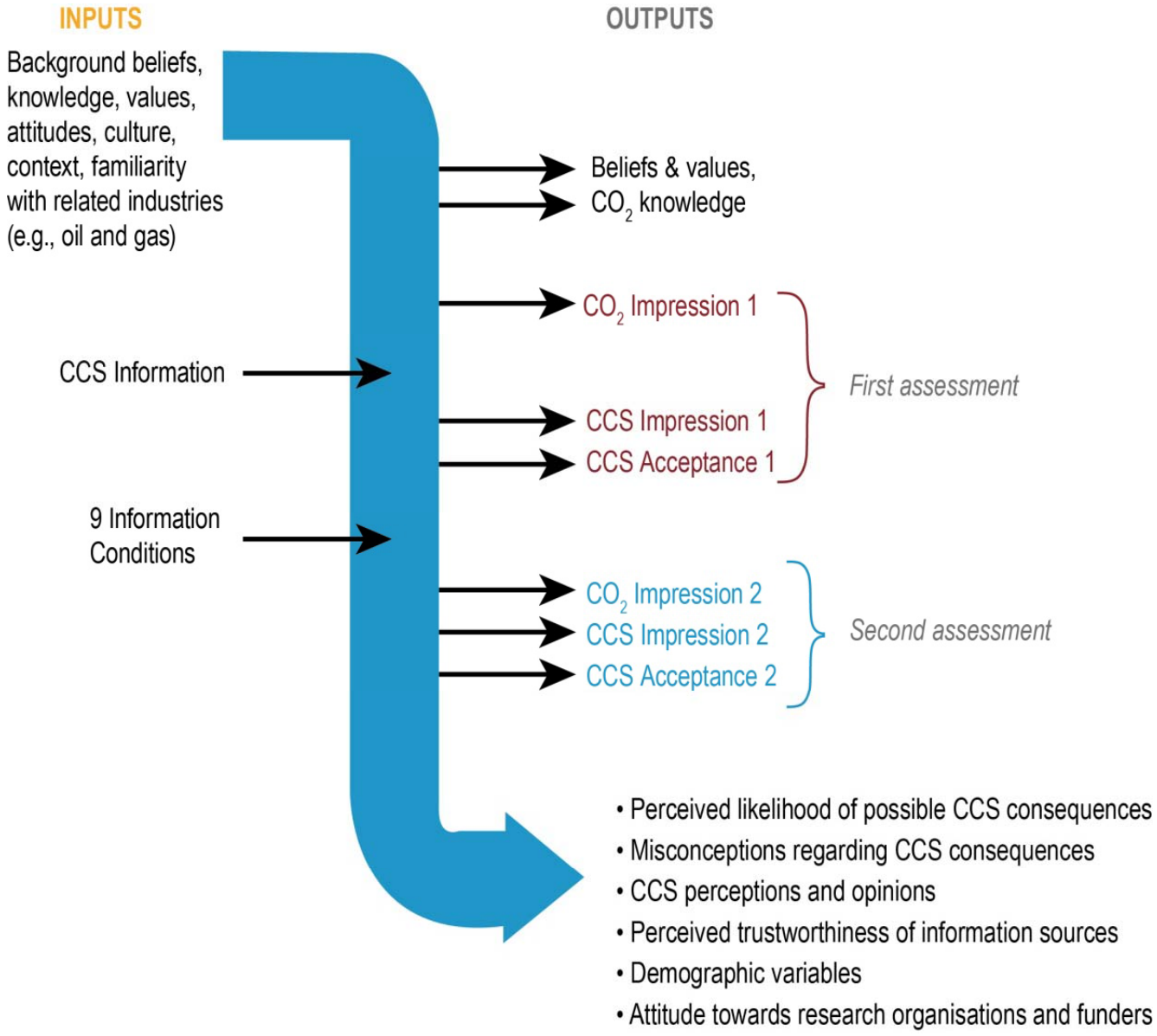
Classification of independent variables

Examples are from a CCS related public perception study "Understanding how individuals perceive carbon dioxide: its relevance to CCS acceptance"

- Exogenous variables:
  - Example: age, education….
- Indigenous variables:
  - Not directly related factors to dependent variable to be explained
    - Example: value and beliefs, CO2 knowledge…..
  - Directly related factors to dependent variable to be explained
    - Example: CCS knowledge, CCS perception…..

Change of R squared

in regression analyses in Understanding how individuals perceive carbon dioxide: its relevance to CCS acceptance”



Change of R squared

in regression analyses in Understanding how individuals perceive carbon dioxide: its relevance to CCS acceptance”

	Independent variable	Country	On shore	Off shore
Opinion 1	Demographics Value and beliefs CO2 knowledge CCS awareness Trustworthy sources	**0.137**	**0.134**	**0.173**
Opinion change (ANOVA)	Only type provided information package	**0.009**	**0.007**	**0.008**
Opinion change (Regression)	Only perception of pieces of information included in provided information package	**0.017**	**0.015**	**0.019**
Opinion 2	Full set (not including CCS impression variables)	**0.440**	**0.361**	**0.424**

- In the case we add CCS impression variables (positivity cleanness, useful ness, safety and maturity), R squared increase more than 0.6 but they tend to hide effects of some of other variables.

Example of literature

- In the case exogenous variables are mainly used:

Case	Number of Observations	R^2	Significant Variables	Direction
Control: Natural Gas Pipeline	827	0.22		
			Gender: Female	Negative
			Race: Asian	Negative
			Registered to Vote	Negative
			Considers Job Important	Negative
			Gun Ownership	Positive
			Union Membership	Positive
			Anti-Environmental Attitudes	Positive
			South Atlantic States [§]	Positive
			West South Central States	Positive
			Mid-East North Central States	Positive
			Mid-West North Central States	Positive
Case	Number of Observations	R^2	Significant Variables	Direction
Compensation: Natural Gas Pipeline	912	0.12		
			Registered to Vote	Negative
			Gender: Female	Negative
			Anti-Environmental Attitudes	Positive
			Ideology: Conservative	Positive
			Mid-East North Central States	Positive
			West South Central States	Positive
			Shop at Wal-Mart	Positive
			West Mountain Region	Positive
			South Atlantic States	Positive
			Mid-West North Central States	Positive

Table 5-2: Regression Analysis of Public Acceptance of a Natural Gas Pipeline

literature cited: Geologic Storage of Carbon Dioxide: Risk Analyses and Implications for Public Acceptance
byGregory R. Singleton B.S., Systems Engineering, University of Virginia, 2002

Example of literature

- In the case indigenous variables are mainly used:

TABLE 1. Factors and Their Coefficients of Risk Perception in the Regression Model[d]

	B	SE B	β	t
constant	0	0.027		0
socioeconomics	0.511	0.038	0.414[c]	13.606
leakage	0.374	0.056	0.230[c]	6.62
pressurization	0.315	0.052	0.188[c]	6.071
diffuse impact	0.124	0.043	0.094[b]	2.849
storage mechanisms	−0.131	0.045	−0.086[b]	−2.925
climate change awareness	−0.105	0.036	−0.084[b]	−2.957
CO_2 knowledge	−0.116	0.046	−0.073[a]	−2.521

[a] $p < 0.05$. [b] $p < 0.01$. [c] $p < 0.001$. [d] $R^2 = 0.52$.

TABLE 2. Factors and Their Coefficients of Benefit Perception in the Regression Model[a,d]

	B	SE B	β	t
constant	0	0.032		0
socioeconomics	−0.520	0.044	−0.421[c]	−11.701
storage mechanisms	0.241	0.053	0.158[c]	4.548
leakage	−0.249	0.067	−0.153[c]	−3.726
climate change awareness	0.190	0.042	0.151[c]	4.506
CO_2 knowledge	−0.145	0.055	−0.091[b]	−2.66
diffuse impact	−0.031	0.051	−0.024	−0.608
pressurization	0.013	0.061	0.008	0.214

[a] $p < 0.05$. [b] $p < 0.01$. [c] $p < 0.001$. [d] $R^2 = 0.33$.

literature cited:Impact of Knowledge and Misconceptions on Benefit and Risk Perception of CCS L Wallquist, VHM Visschers and M Siegrist - Environ. Sci. Technol., 2010

Example of literature

- A researcher's comment:

…….in many social science settings, an Rsquare of 9% is considered respectable. That's about as good as it gets in most psychology studies where two distinct variables are correlated with each other. Example: extraversion explains only about that much of the variation is sales effectiveness.

When you measure variables with error, that can lower your Rsquare. What is S&P500 taken to be a measure of? The overall health of the economy, or just the stock market? Or just itself?

http://www.marketingprofs.com/ea/qst_question.asp?qstID=21047

Example of factors to influence of fitness

- Accuracy of measurements (size of error)
- Resolution of measurements
- Less number of unobserved factors
- Strength of causality
- Fundamental randomness
- …

Conclusion

- In social science, to examine the effectiveness of a factor, the size of R squared does not matter.
- However, we need to explain why the R squared is low if it is the case.
- At least, we explain potential important covariates (independent variables) are included or not .
  - If those covariates are included in the model and the R squared is still low, we would claim the measurement of dependent variable and some independent variables are not accurate.
  - If those covariates are not included, we should mention why we cannot include those and claim

    “further research is necessary!”
- High R squared would sometime be dangerous when we use directly related indigenous factors to dependent variable. They might hide effects of some of other variables.
- Maybe better conduct SEM (path anaylsis)

Conclusion 2

In social science setting,

- List all potentially influential factors
- Check simple correlation
- Conduct multiple regression
- Check residual (linearity)
- Again try to find hidden factors
- If the list of variables for input of regression is defendable and there is not much multi-colinearity, the model is considered to be fine even with low R-squared.
- Maybe better conduct SEM (path anaylsis)

Thank you!

Contact: kenshi.itaoka@mizuho-ir.co.jp

Backup

Example of literature

- Analysis

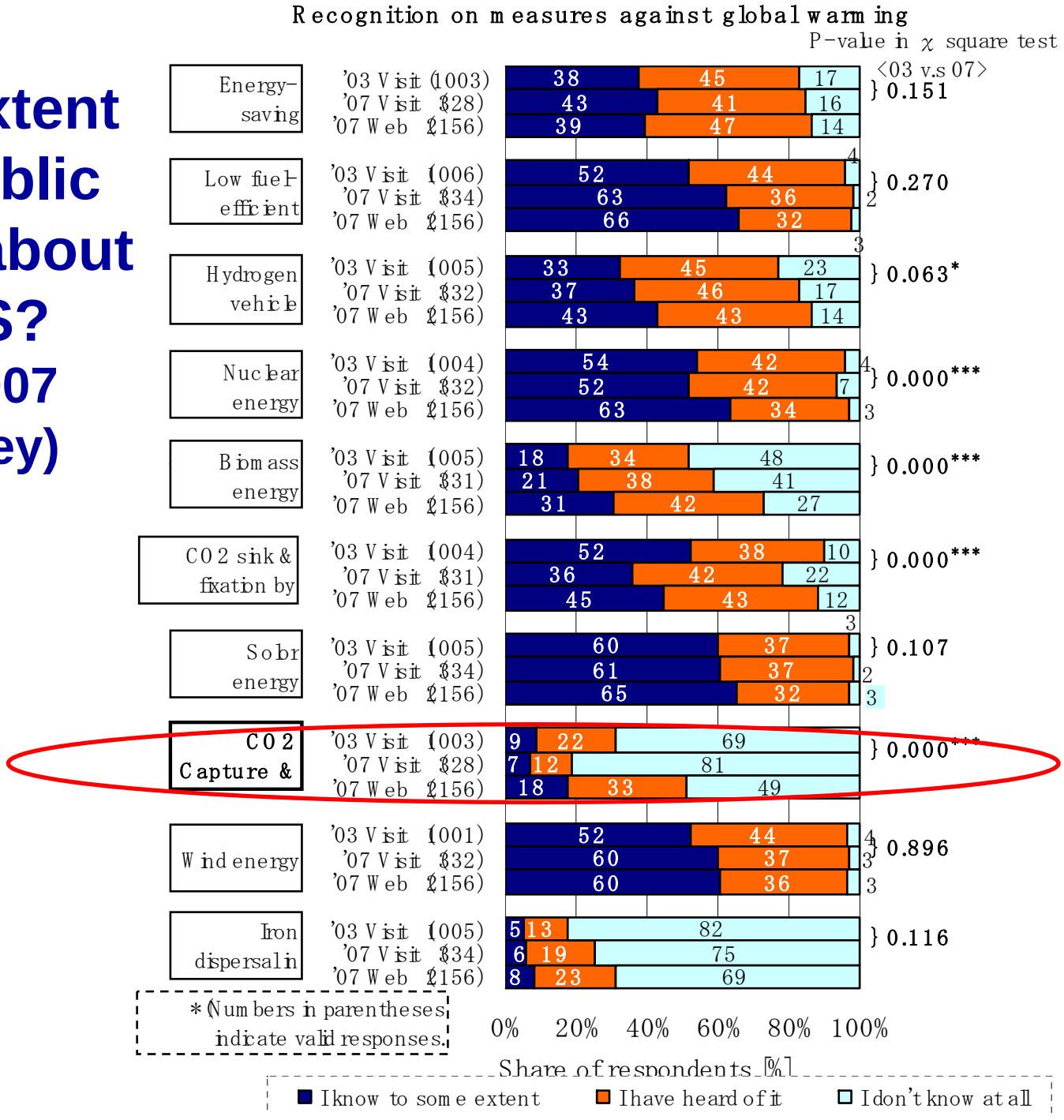
Table 6
Independent variables and the regression model.

Variables	Descriptions	Coefficients	*T*-Statistics
Age	Respondents directly asked to report their age (mean = 33.52, SD = 11.08)	−.004	−1.377
Gender	Male = 1, female = 0	−.045	−.751
Education	4 categories (primary school = 1, middle school = 2, high school = 3, college and above = 4)	−.016	−.207
Income	1–6 scale (1 = less than 20 k/yr, 6 = 60 k/yr+, intermediate gradations)	.015	.732
Satisfaction	Respondents satisfied with the current life (1–10 scale, unsatisfied at all/very satisfied, mean = 6.8)	.044	2.640[a]
Environment	Respondents' evaluation on China's environmental condition (1–6 scale, very bad/very good)	−.004	−.156
Pathways	China's development pathways (economic priority = 1, environmental priority = 2, equally important for economic growth and environmental protection = 3)	−.116	−2.157[b]
Climate change	Awareness of climate change causes, impacts and solutions (aggregated data calculated from the 15 statements)	.036	.670
CCS knowledge	Know about it = 3, heard about it = 2, never heard about it = 1	.093	1.675[c]
Maturity of CCS	7-point scale from strongly disagree to strongly agree	−.039	−1.767[c]
Uncertainties and risks	7-point scale from strongly disagree to strongly agree	−.080	−3.498[a]
CO_2 emission reductions	7-point scale from strongly disagree to strongly agree	.303	13.532[a]
Government CCS policy	7-point scale from strongly disagree to strongly agree	.099	4.262[a]
N	487		
R^2	0.663		
F(13, 487)	28.565[a]		

[a] Significant at $p < 0.01$.
[b] Significant at $p < 0.05$.
[c] Significant at $p < 0.1$.

literature cited:The public perspective of carbon capture and storage for CO_2 emission reductions in China
H Duan - Energy Policy, 2010

To what extent do public know about CCS? (2007 survey)



Background & objective

Background

- Recently implementation of large demonstration projects and commercial projects has become an important agenda for GHG mitigation in the world.
- In this move, the issue in public acceptance of CCS expands to cover from policy formulation in national policy arena to project implementation sites in local policy arena.
- Therefore needs for assessment of public opinions on CCS have changed to include not only those for general public in national policy context but also those for local public in project implementation context.