rego

Stata module for decomposing goodness of fit according to Shapley and Owen values

Frank Huettner and **Marco Sunder**
Department of Economics
University of Leipzig, Germany

Presentation at the UK Stata Users Group meeting, London, 13/14 Sept. 2012

Motivation

- Challenges in interpreting regression results by the analyst (the quick and dirty way)
  - Sign *econometrics*
  - Confusion of statistical significance and *economic* significance
  - Interaction terms

- Challenges to readers
  - Units of measurement
  - "Omitted coefficients"

Decompose

$$\mathbb{R}^2$$

Motivation

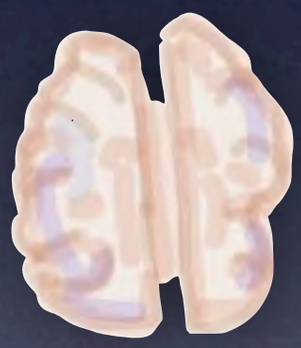
- Imagine regressor variables as "players"
  - Cooperative game theory: games with transferable utility
  - Regressors may form coalition – how to distribute the gains from cooperation?
  - **Shapley** (1953) and **Owen** (1977) **Values** as means to decompose goodness of fit
  - Axioms under which these values are unique solutions
- "**hierarchical partitioning**"
  (Lindeman et al. 1980, Chevan & Sutherland 1991)
- Previous implementations:
  Stata: *shapley* (Kolenikov 2000)
  R: *relaimpo package* (Grömping 2006)

Illustrative calculation
of the Shapley Value

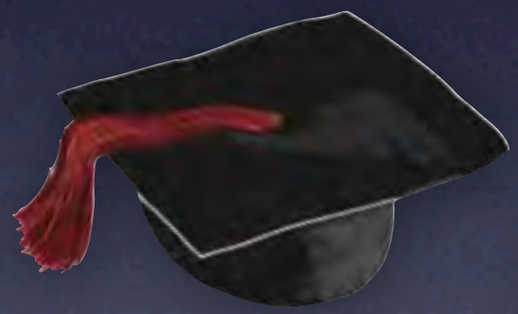
- Assume the "full model" is…

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

log wage



intelligence



schooling

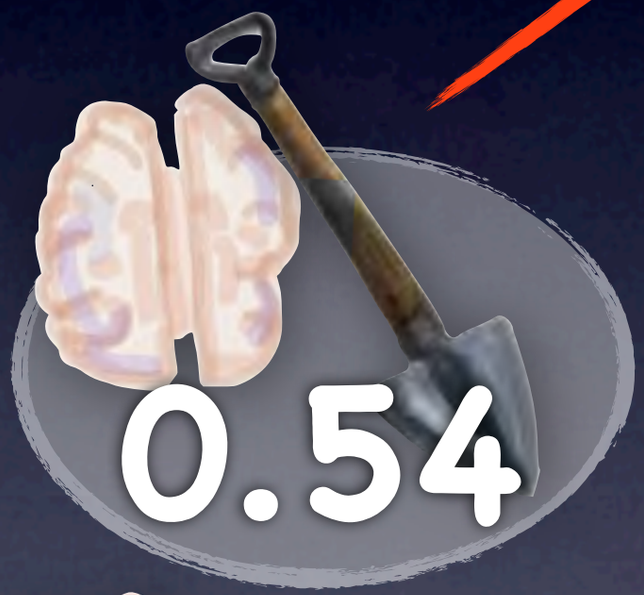
experience

Worths in 2^k (sub-)models

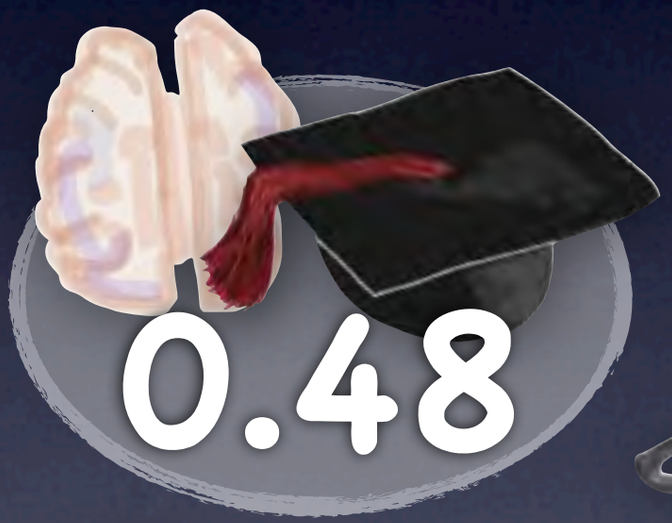
0.72

0.00

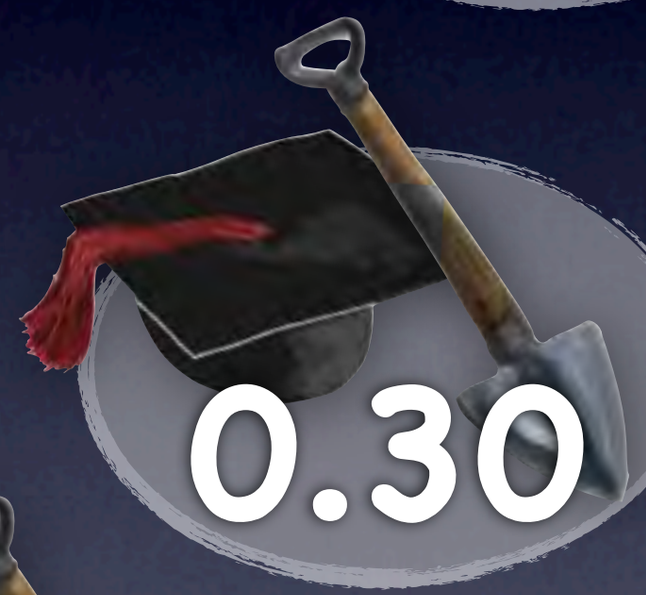
Δ = 0.18



0.54

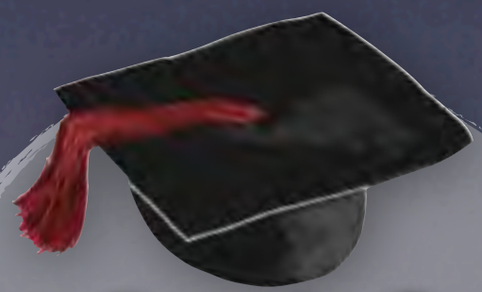


0.48

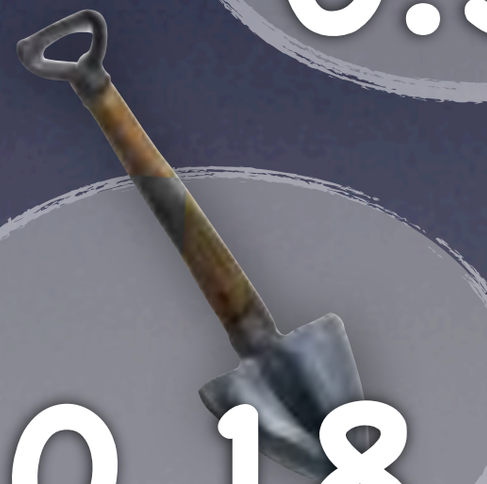


0.30

0.24



0.12



0.18

Marginal contributions (MC)

remove first → remove first → remove first

0.72	0.30	0.18	0.00
0.72	0.30	0.12	0.00
0.72	0.54	0.18	0.00
0.72	0.54	0.24	0.00
0.72	0.48	0.12	0.00
0.72	0.48	0.24	0.00

$$\mathrm{Sh}(x_j) = \frac{1}{k!} \sum_{\text{all permutations}} MC$$

$$\mathrm{Sh}(\text{  }) = \frac{1}{6} \cdot (0.12 + 0.12 + 0.18 + 0.18 + 0.12 + 0.24) = 0.16$$

$$\mathrm{Sh}(\text{  }) = \frac{1}{6} \cdot (0.42 + 0.42 + 0.36 + 0.24 + 0.36 + 0.24) = 0.34$$

$$\mathrm{Sh}(\text{  }) = \frac{1}{6} \cdot (0.18 + 0.18 + 0.18 + 0.30 + 0.24 + 0.24) = 0.22$$

$$\boldsymbol{\Sigma} = 0.72$$

Axioms

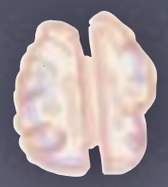
1) **Efficiency**
GOF of full model is decomposed among the regressors

2) **Monotonicity**
Increase in R^2 must not decrease the value

3) **Equal treatment**
Perfect substitutes (in terms of GOF) receive the same value

The **Shapley Value** is the **only** value that satisfies these properties (Young 1985).

A priori grouped regressors

- Analyst may believe that some regressor variables belong together, e.g.: polynomial terms, region dummies
- Regressor variables are partitioned: $\mathcal{G} = \{G_1, ..., G_\ell, ..., G_\gamma\}$
- When thinking about marginal contributions of variables from G_ℓ, variables of $G_{q \neq \ell}$ must be completely absent or present
- This limits the set of admissible model permutations
  ☞ **Owen Value**
- E.g.:  "nature",  "nurture"

4 permutations respect $\mathcal{G}$

0.72	0.30	0.18	0.00
0.72	0.30	0.12	0.00
0.72	~~0.54~~	~~0.18~~	0.00
0.72	0.54	0.24	0.00
0.72	~~0.48~~	~~0.12~~	0.00
0.72	0.48	0.24	0.00

Ow( ) = 0.33 Ow( ) = 0.165 Ow( ) = 0.225

Axioms

1) **Efficiency***
2) **Monotonicity***
3) **Equal treatment of players**
4) **Equal treatment of groups**

The **Owen Value** is the **only** value that satisfies these properties (Khmelnitskaya and Yanovskaya 2007).

Stata implementation

- With grouping, "large" models become possible
- User decides for which groups to calculate Owen values
- R^2 calculated from covariance structure of the data
- Syntax uses "\" to indicate group boundaries in *varlist*
- Computation in Mata
- Bootstrapping option

Viewer (#1) [help rego]

help rego

Advice Contents What's New News

help rego

Title

rego — Linear regression with Shapley and Owen decomposition of R-squared (for Stata 9+)

Syntax

rego *varlist* [*if*] [*in*] [, *options*]

options	description
Main	
detail	calculate Owen values for individual variables
vce(*vcetype*)	determine what standard errors are computed: ols, robust, cluster *clustervar*.
noperc	do not display Owen/Shapley values as percentages of overall R-squared
force	perform calculations even if it would take a lot of time
Bootstrap	
bsreps	number of bootstrap replications
level	confidence level, in %

xi (Stata 9+) and time series operators (Stata 11+) may be used, but factor variables are not supported.

Description

rego uses results from regress and decomposes the share of explained variance (measured by R-squared) into contributions by indvidual regressor variables or groups of regressor variables. The former case is calculated as Owen value, the latter as Shapley value (for groups). In the special case that each group consists of only one regressor, both quantities coincide. If not suppressed by the noperc option, the contributions are displayed as percentages of overall R-squared.

Remarks

Groups of regressors are specified with the help of the backslash ("\") symbol. It must be placed in the varlist after the last variable of each group. If no groups are defined, rego assumes that each regressor belongs to a group of its own and calculates the "traditional" Shapley value.

Instead of using the detail option after the comma, the user may also specify "(detail)" (without quotation marks)

Wage regression: German male employees

OLS regression results with decomposition of R^2 (in %)

			R^2 decomposition (%)	
Group	Regressor	Coef.	Owen	Group
1	SCT	0.789 *	3.0	33.2
	SCT × EDUC	−0.048 *	8.3	
	EDUC	0.103 ***	21.9	
2	EXPER	0.025 ***	7.0	11.0
	$(\text{EXPER})^2/100$	−0.041 ***	4.0	
3	TENURE	0.017 ***	9.3	14.3
	$(\text{TENURE})^2/100$	−0.029 **	5.0	
4	MARRIED	0.084 ***	5.0	5.0
5	Firm size	(3 dummies) ***		14.7
6	Industry	(6 dummies) ***		5.5
7	Region	(14 dummies) ***		16.2
Observations		850		
Full model R^2		0.501		

Remark: */**/*** denotes statistical significance at the 10% / 5% / 1% level for individual variables (t-test) or groups of dummy variables (F-test), based on the heteroscedasticity-robust covariance matrix.

Wage regression: German male employees

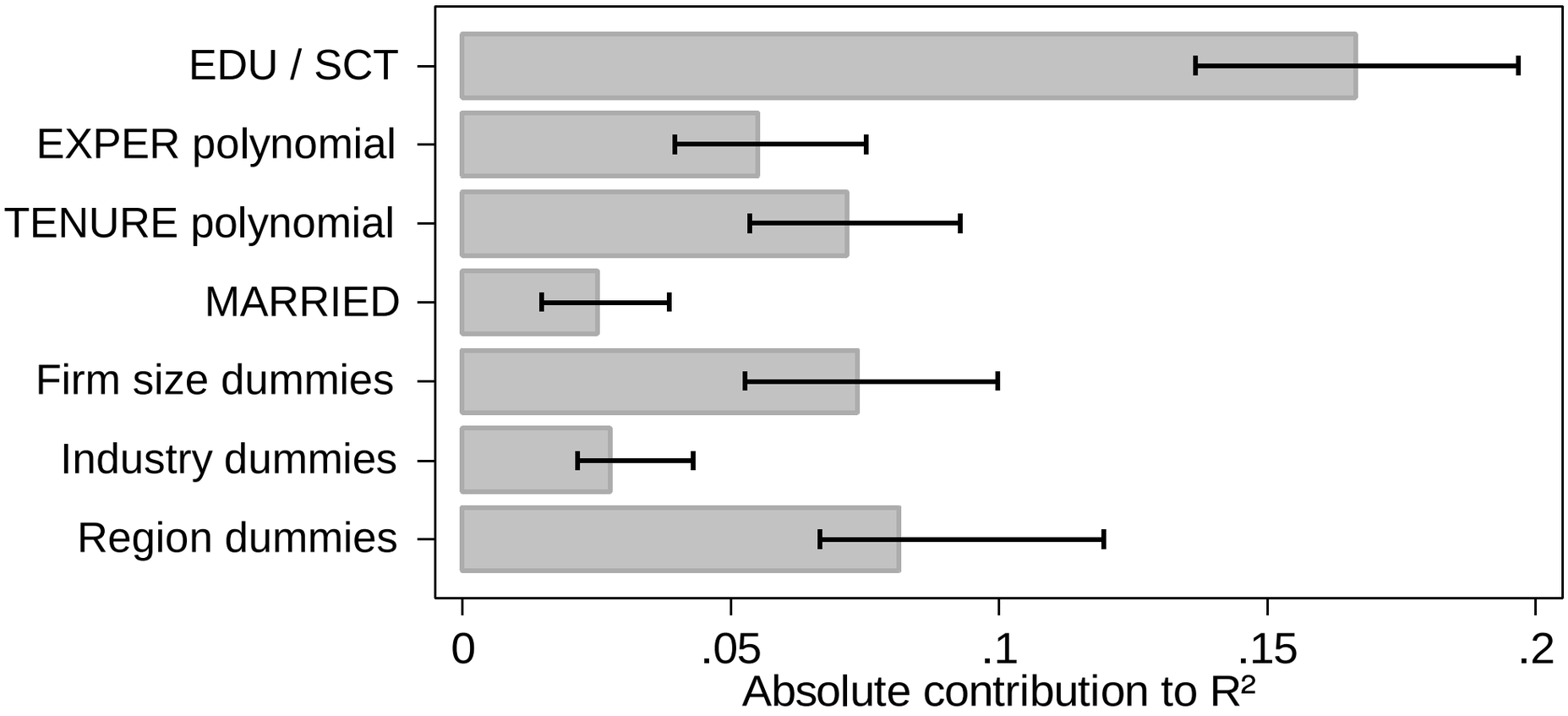


FIG 1. *Decomposition results for groups, with 90% bootstrap confidence intervals, based on 5000 bootstrap replications.*

Wage regression: German male employees

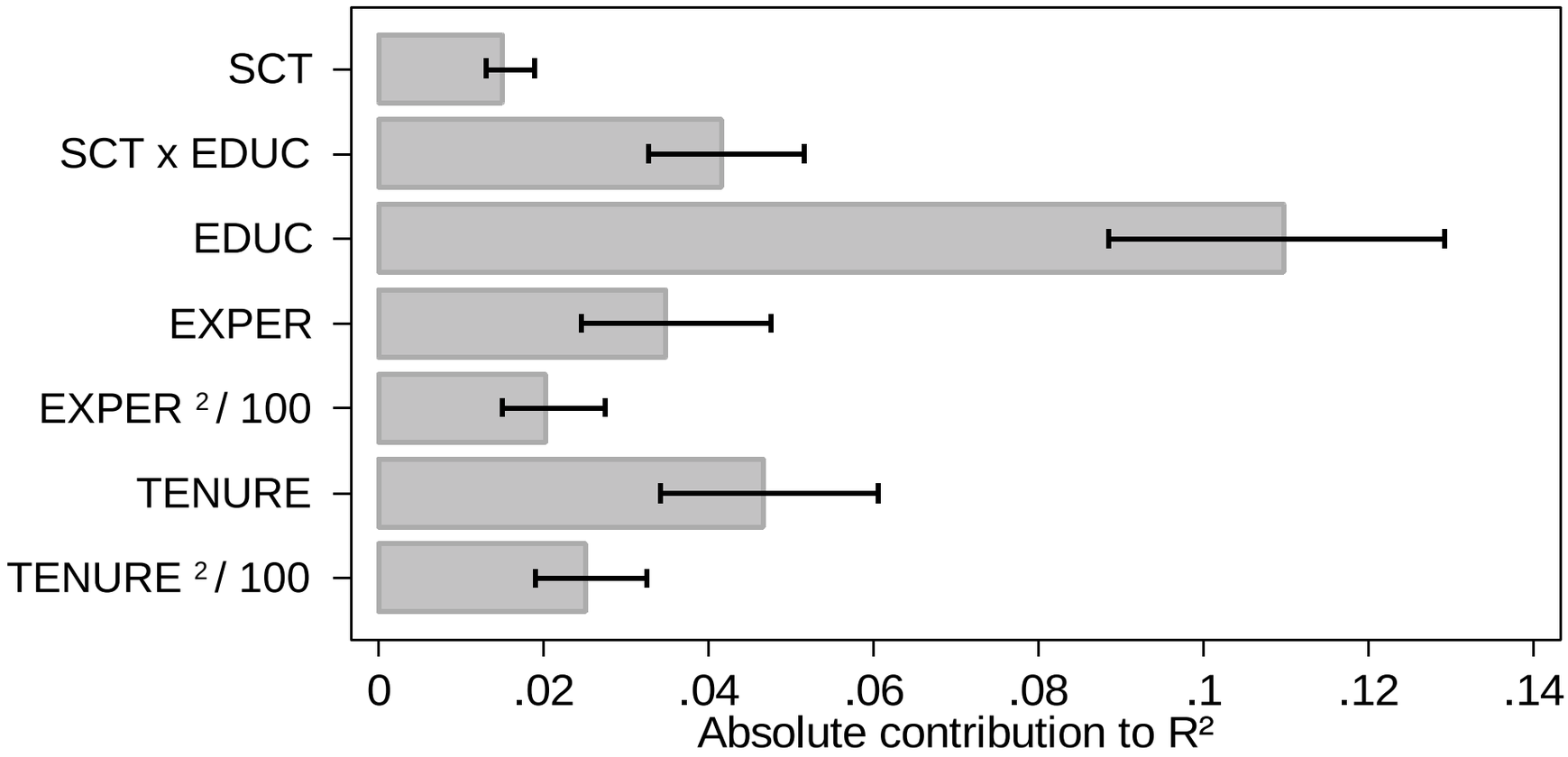


FIG 2. *Owen value decomposition results for 'human capital' variables, with 90% bootstrap confidence intervals, based on 5000 bootstrap replications.*

Limits

- Only OLS with R^2 decomposition at this time
- Does not yet accept factor variables or weights
- …

Possible extensions?

- Decomposition of other measures (e.g., AIC)
- More levels of aggregation
- Essential regressors / fixed effects model (*econ*)

Thank you!

www.uni-leipzig.de/~rego


Electronic Journal of Statistics
Vol. 6 (2012) 1238–1250
ISSN: 1935-7524
DOI: 10.1214/12-EJS710

Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values

Frank Huettner and Marco Sunder

Faculty of Economics and Management Science, University of Leipzig, Grimmaische Str. 12, 04109 Leipzig, Germany
e-mail: huettner@wifa.uni-leipzig.de; sunder@wifa.uni-leipzig.de

Abstract: We advocate the decomposition of goodness of fit (gof) contributions of (groups of) regressor variables according to the Shapley value or—if regressors are exogenously grouped—the Owen value because of the attractive axioms associated with these values. A wage regression model with German data illustrates the method.

AMS 2000 subject classifications: 62J05, 62P20, 91A12.
Keywords and phrases: Shapley value, Owen value, variance decomposition, regression games, GSOEP.

Received January 2012.

1. Introduction

One of the unwritten conventions in applied econometrics is that authors provide their readers with some goodness-of-fit measure (GOF) at the end of each regression table. Very rarely, however, the GOF is allocated to individual regressor variables, even though—or because—the literature provides numerous different approaches to do so [3, 9]. Rather, the discussion of 'relevance' of regressor variables is often confined to the sign and p-value of their corresponding coefficients [12]. Due to space constraints, many coefficients are not even reported, leaving readers bewildered as to how important (with respect to GOF) such 'omitted' variables were compared to those variables of primary interest.

In the present paper, we advocate the method that employs the Shapley value [17]—and its generalizations—to distribute the GOF of the model among the regressor variables, henceforth Shapley value decomposition [20]. This method takes account of the interplay of regressor variables in sub-models and is calculated on the basis of information on the same type of GOF in these sub-models. Its attractiveness also stems from the fact that it emerges as the unique solution to the decomposition problem under a sound set of assumptions.

A generalization of the Shapley value, the Owen value [14], allows for decomposition in the context of exogenously grouped regressors as is suggested by Stavrocks [18]. Such groups may arise, e.g., if the model includes polynomial terms of a variable, dummy variables that recode a categorical variable, or variables that are conceptually related for other reasons. Under such circumstances,

Huettner & Sunder (2012)
Electronic Journal of Statistics