S-values: Conventional measures of the sturdiness of the signs regression coefficients

Ed Leamer

December 2015

BITSS: Berkeley

Related Paper

S-values and Bayesian Weighted All-Subsets Regressions

Introductory Comments

- How I started on this journey and where I am now.
- Specification Searches
  - Hypothesis Testing Searches
  - Interpretative Searches
  - Proxy Variable Searches
  - Simplification Searches
  - Data Selection Searches
  - Data-instigated Models
- Transparency alone is not enough. We need accountability.
  - If you tried only one regression and reported it, but I found the very same regression after 1000 trials and reported it and also reported the 999 others, is there a difference in the reliability of my final equation and your one-and-only-one equation?? It's the same regression after all.

Two Sources of Model Error

- t-values measure the impact of ***statistical uncertainty*** on the signs of coefficients.
- s-values measure the impact of ***model ambiguity*** on the signs of coefficients. (s stands for "sturdy.")

t-values: An Unfamiliar Measure of Model Ambiguity

Theorem: When a variable is omitted from a linear regression, an estimated coefficient cannot change it's sign if the coefficient of the retained variable has a t-value that exceeds in absolute value the t-value of the coefficient of the omitted variable.

Leamer, "A Result on the Sign of Restricted Least Squares Estimates," *Journal of Econometrics*, 3 (1975), 387-390.

Operational Significance:
If you want to change the sign of a coefficient, omit variables with bigger t-values!

Big t good, Small t bad.
Is That All That Can be Said??

- **Statistical Uncertainty:** Compare the absolute t-value with the number 2.
- **Model Ambiguity:** Compare the absolute t-value with the absolute t-values of other coefficients.

Two Variable Case: Is the Sign of a Regression Coefficient Known?

Statistical Uncertainty: How much does knowledge about the sign depend on the limited sample size and the degree of collinearity

Model Ambiguity: How much does knowledge about the sign depend on the choice of explanatory variables?

Notation

- Standardize the data to have zero means and unit variances.
- The data consists of a sample size and three correlations

  r_1=correlation(y,x_1), r_2=correlation(y,x_2)

  ρ = correlation(x_1,x_2)
- OLS estimates

  $$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} r_1 - \rho r_2 \\ r_2 - \rho r_1 \end{bmatrix} / (1 - \rho^2)$$

2-Variable Sturdiness Theorem

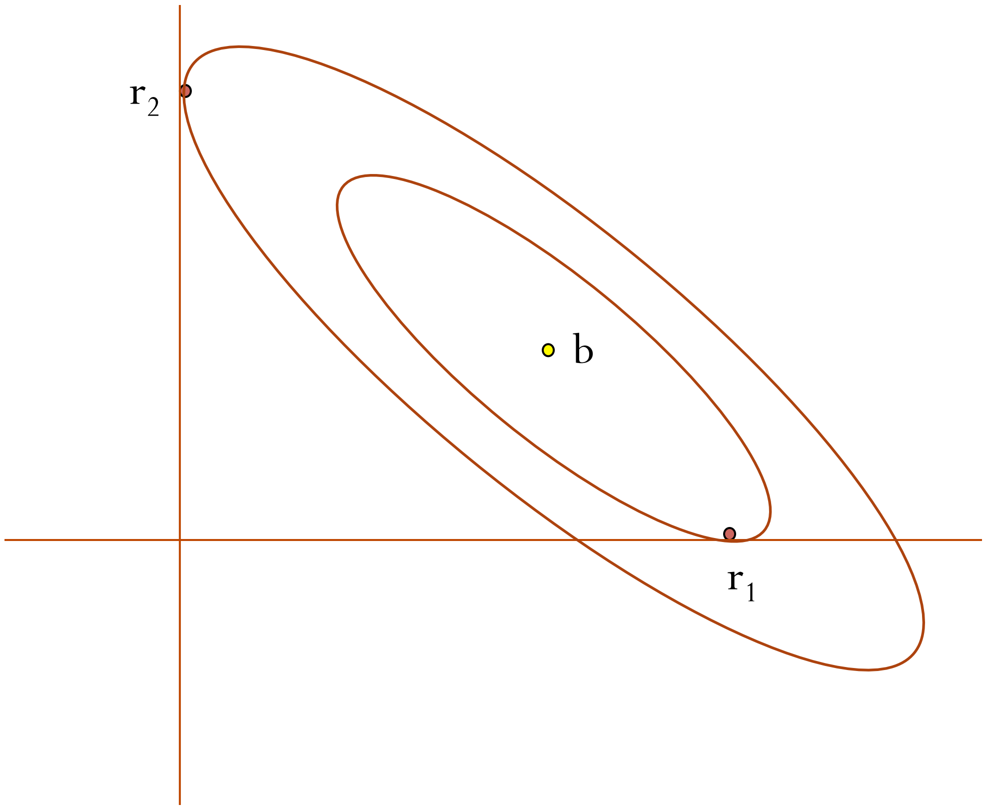
- Theorem
  - A necessary and sufficient condition for b_1 and r_1 to have the same sign is $r_2/r_1 \, \rho < 1$.
  - A necessary and sufficient condition for b_2 and r_2 to have the same sign is $r_1/r_2 \, \rho < 1$.
  - A necessary and sufficient condition for the signs of both coefficients to be unchanged when a variable is omitted is $\max(r_1/r_2, r_2/r_1) \, \rho < 1$.
  - Note $r_2/r_1 = t_2/t_1$.
- Implications
  - A value of ρ equal to zero makes sturdy estimates.
  - The sign of ρ opposite r_2/r_1 makes sturdy estimates.
  - Equal values of r_1 and r_2 make a sturdy estimates.
  - t-values rank the sturdiness of coefficients.

No sign change: The sign of ρ opposite r_2/r_1 .

Likelihood ellipse points away from the origin.

Variables are substitutes: sum of the coefficients is well estimated.

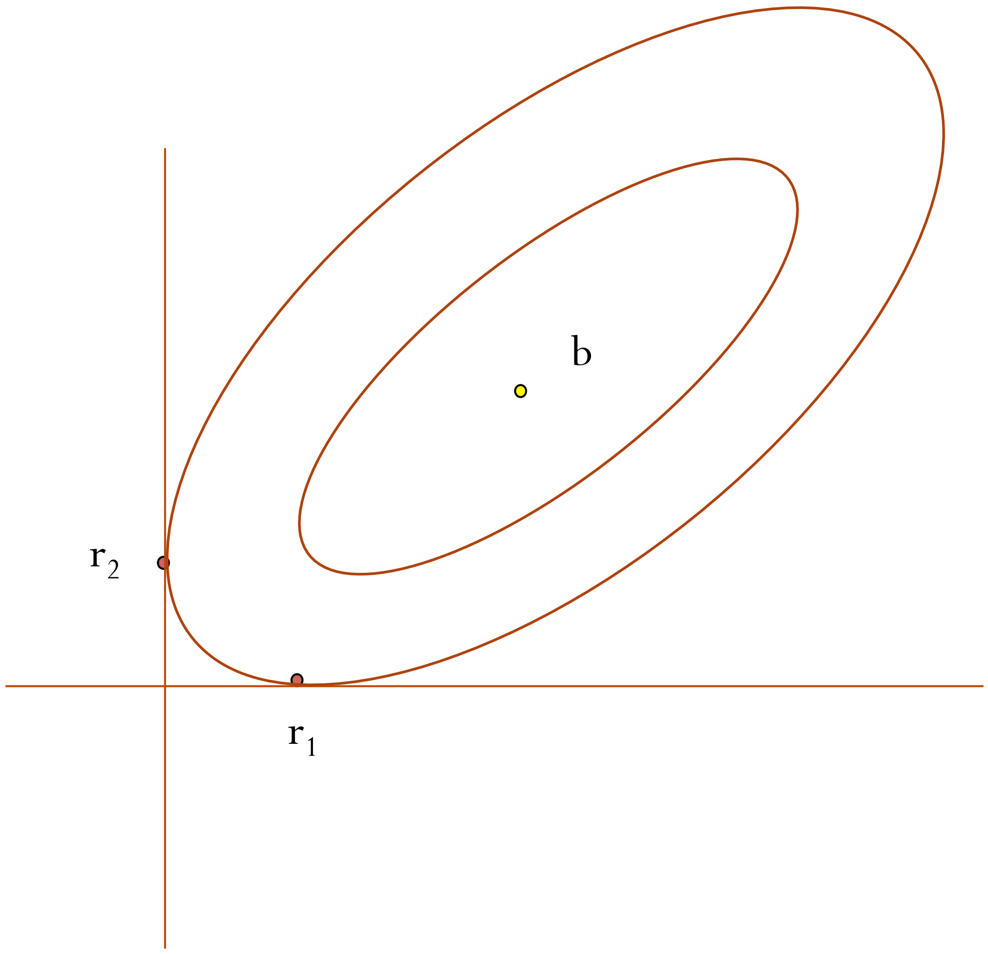
Shrink one but enlarge the other.



No sign change: The sign of ρ same as r_2/r_1 but r1=r2

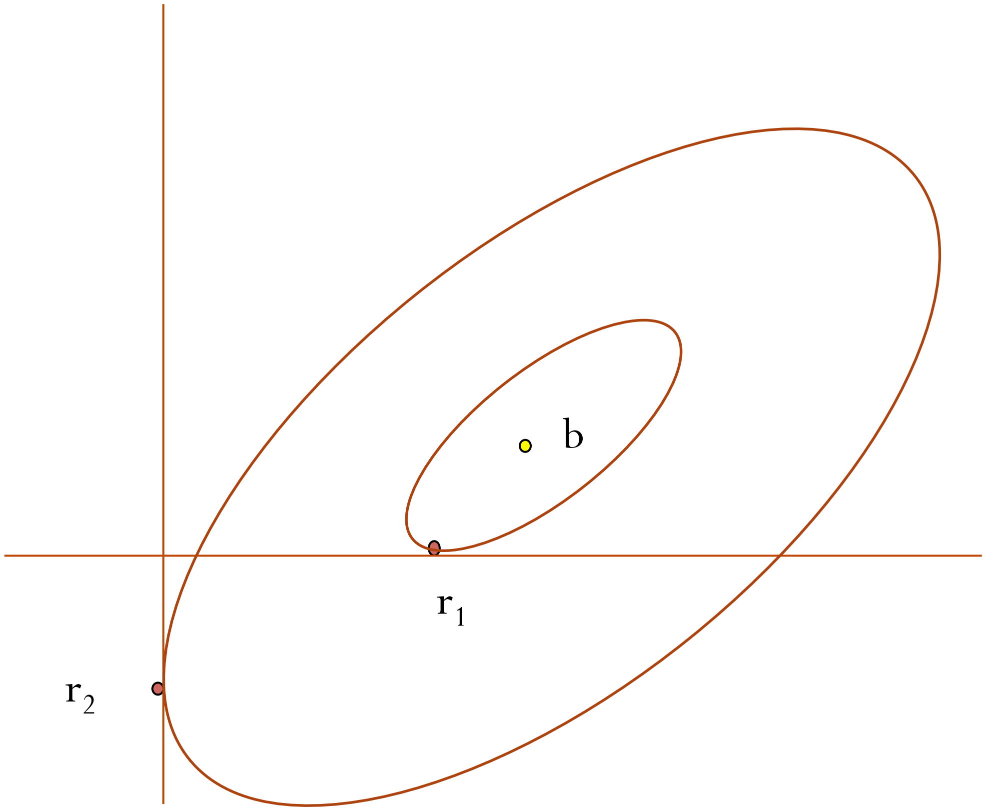
Likelihood ellipse points toward the origin: Shrink together.

Variables are complements: difference of the coefficients is well estimated.



<u>Sign Change</u>: The sign of ρ same as r_2/r_1 and $|r_1|>|r_2|$

Likelihood ellipse points to the side of the origin favoring one coefficient over the other.
Shrink one coefficient more than the other



An s-value: measures model ambiguity the way a t-value measures statistical uncertainty

- t-value = estimate / half the length of the confidence interval
- Center and spread of ambiguity interval
  - Center is average of $(r_1 - \rho r_2)/(1-\rho^2)$ and r_1 :
    $[r_1 - \rho r_2 + (1-\rho^2) r_1] / 2(1-\rho^2)$
  - Distance between the two estimates is absolute value of
    $(r_1 - \rho r_2)/(1-\rho^2) - r_1 = [r_1 - \rho r_2 - (1-\rho^2) r_1] / (1-\rho^2)$
- s-value is the average divided by half the distance:
  $$s_1 = [(2-\rho^2) r_1 - \rho r_2] / \text{abs}[\rho^2 r_1 - \rho r_2]$$

Multivariate Regression with Hard Constraints??

- Conjecture: The alternative signs of the regression coefficients in the 2^k regressions formed by omitting subsets of the k explanatory variables can be found by computing the unconstrained regression and the k regressions with one-at-a-time omissions of the k explanatory variables.

Multivariate Regression With “Soft” Constraints

Regression Setting

- The nx1 vector $\mathbf{y}$ is normally distributed with mean $\mathbf{X\beta}$ and covariance matrix $\sigma^2\mathbf{I}$, where $\mathbf{X}$ is an nxk matrix of explanatory variables, $\boldsymbol{\beta}$ is a kx1 vector of unknown coefficients and σ^2 is a scalar variance applicable to all n observations.
- The OLS estimate of $\boldsymbol{\beta}$ is $\mathbf{b} = (\mathbf{X'X})^{-1}\ \mathbf{X'y}$.
- The likelihood function is constant on the ellipsoids $c = (\boldsymbol{\beta}-\mathbf{b})'\mathbf{X'X}(\boldsymbol{\beta}-\mathbf{b})$

Pooling Two Data Sets: Matrix Weighted Averages

- If there are two sets of observations $(\mathbf{y}_1, \mathbf{X}_1, \mathbf{y}_2, \mathbf{X}_2)$ with different residual variances, $\sigma_1^{\ 2}$ and $\sigma_2^{\ 2}$, the data can be transformed into homoscedastic form by dividing by the standard deviations, σ_1 and σ_2 to produce the pooled estimate:

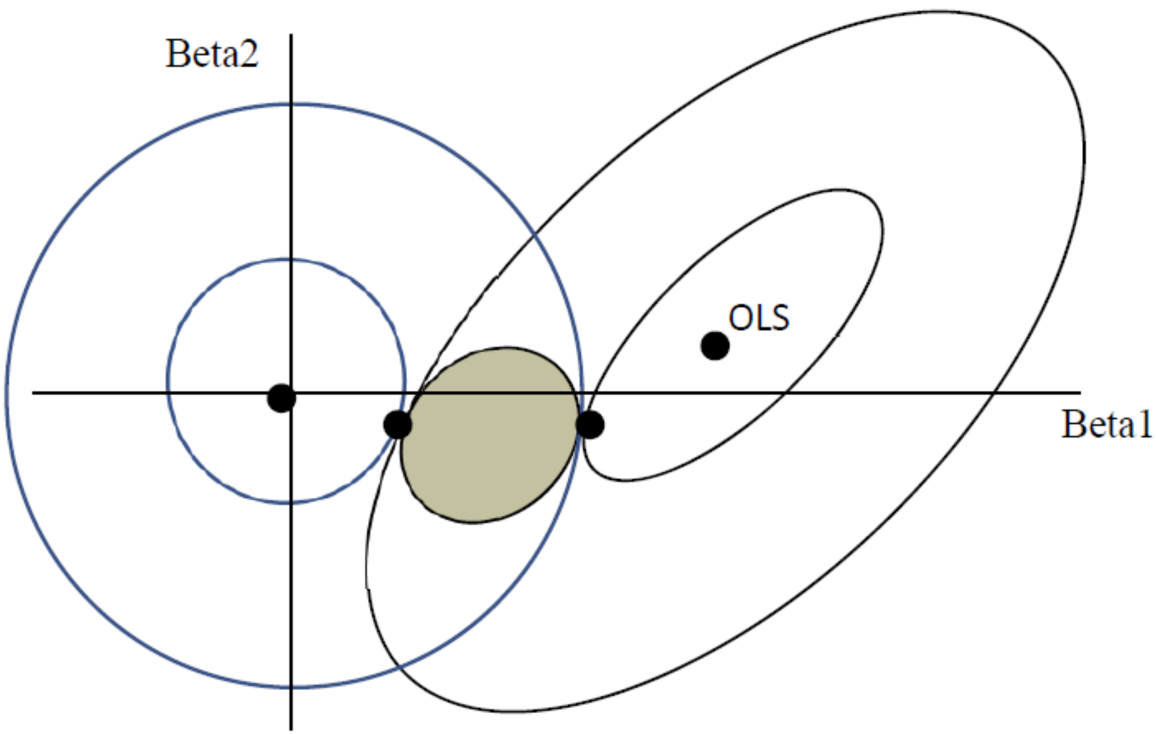
$$\mathbf{b} = (\mathbf{X'X})^{-1}\,\mathbf{X'y} = [\mathbf{X}_1'\,\mathbf{X}_1/\ \sigma_1^{\ 2} + \mathbf{X}_2'\,\mathbf{X}_2/\ \sigma_2^{\ 2}]^{-1}\,[\mathbf{X}_1'\,\mathbf{y}_1/\ \sigma_1^{\ 2} + \mathbf{X}_2'\,\mathbf{y}_2/\ \sigma_2^{\ 2}]$$

$$= [\mathbf{H}_1 + \mathbf{H}_2]^{-1}\,[\mathbf{H}_1\,\mathbf{b}_1 + \mathbf{H}_2\,\mathbf{b}_2]$$

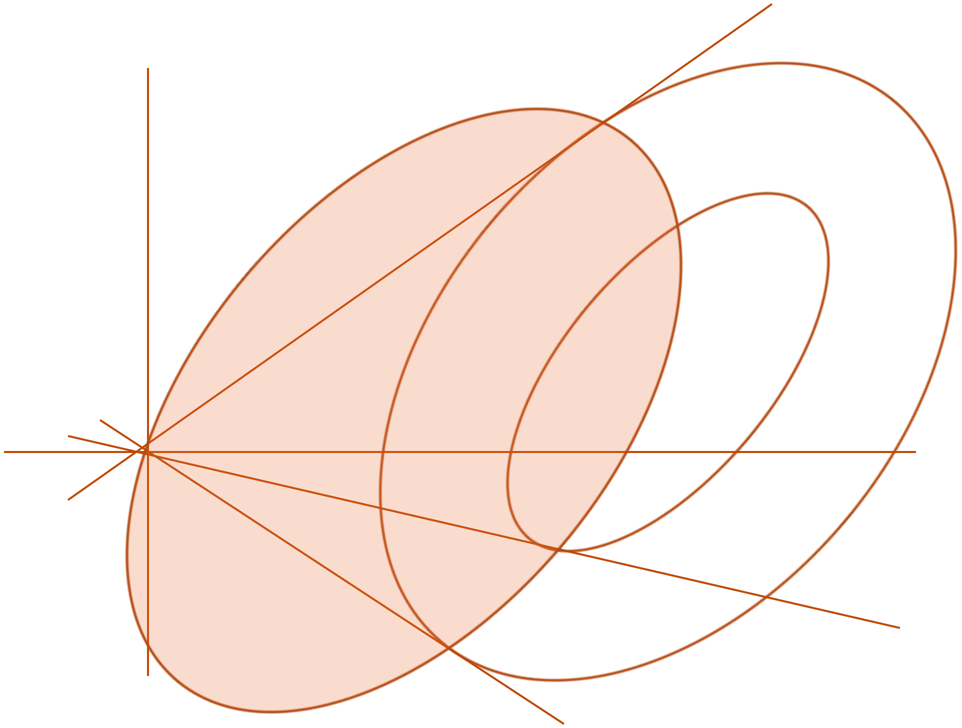
- $\mathbf{H}_i = \mathbf{X}_i'\,\mathbf{X}_i/\ \sigma_i^{2}$, $\mathbf{H}_i\,\mathbf{b}_i = \mathbf{X}_i'\,\mathbf{y}_i/\ \sigma_i^{2}$

Items to Report: Implications of a special fictitious prior data set.

Figure 1 OLS and Three Estimates Closer to Zero

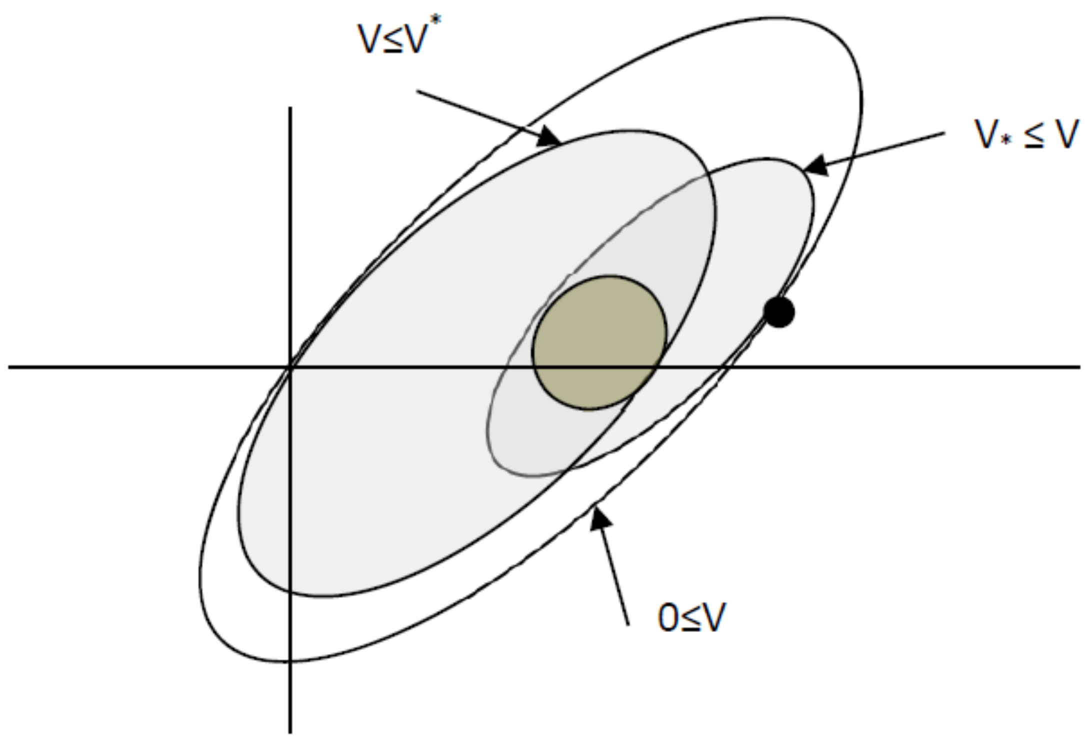


Feasible Ellipse



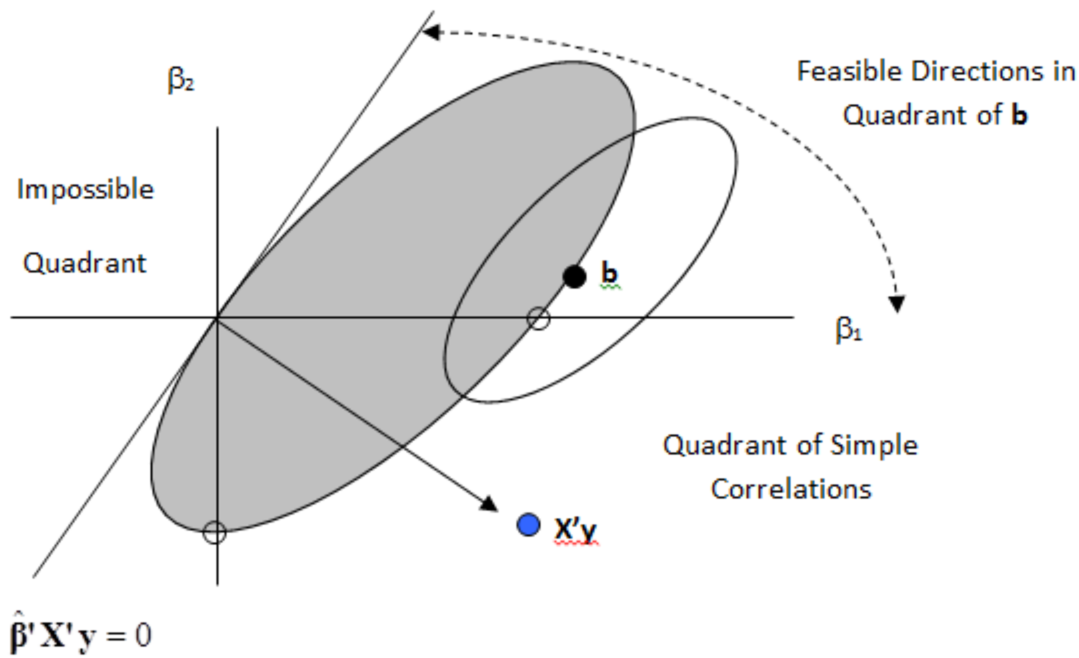
Intervals of Prior Covariance Matrices Map Into Ellipsoids of Estimates

Figure 2 Four Ellipses of Estimates

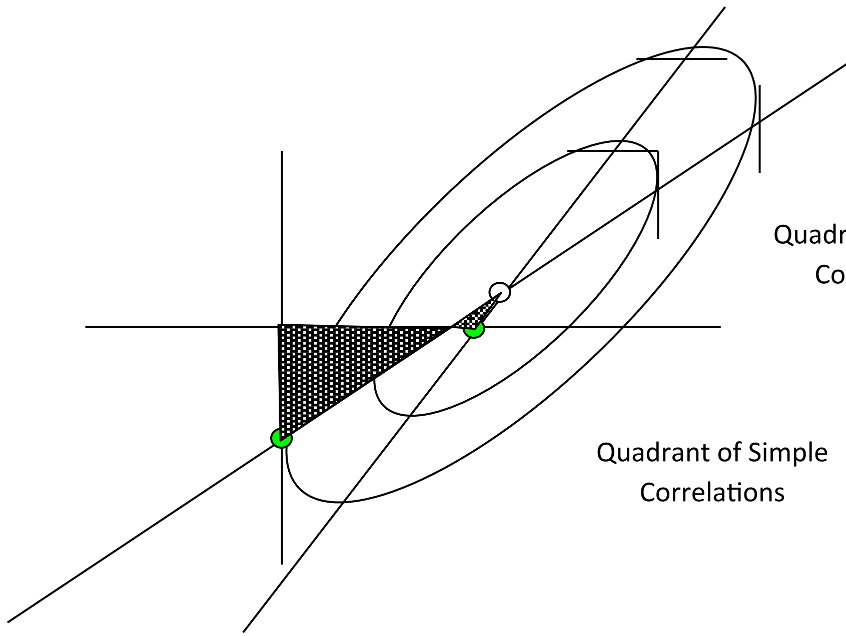


The Orthant of Simple Correlations Is Special

Figure 3 The Feasible Ellipse and the Impossible Orthant



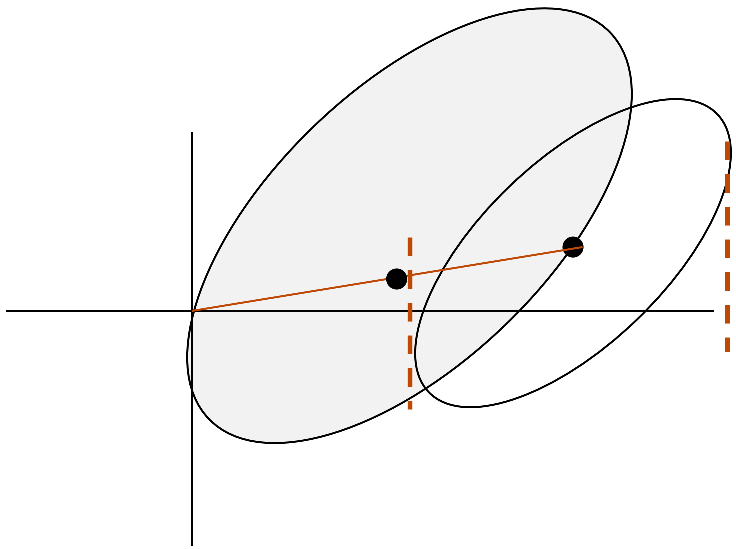
The Orthant of Simple Correlations Is Special When the Prior is Strong and the Prior covariance matrix is diagonal



The Orthant of Partial Correlations Is Also Special

Theorem: If the prior precision matrix is diagonal, then estimates may not lie in opposite orthants.

ALERT:
s-values can be proportional to t-values



Standardized Variables Are Amenable to Context-Independent Conventional Prior Covariance Matrices

$$y_t = \mathbf{x}_t{}'\boldsymbol{\beta} + \varepsilon_t$$

$$\sigma_y^2 = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta} + \sigma_\varepsilon^2$$

$$R^2 = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta}$$

$$Var(\boldsymbol{\beta}) = v^2\boldsymbol{I},$$

$$trace(\boldsymbol{\beta}'\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta}) = trace(\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta}\boldsymbol{\beta}')$$

$$E(R^2) = trace(v^2\boldsymbol{\Sigma}_{xx}) = v^2 k$$

$$v_L^2 = \frac{\min \mathrm{E}(R^2)}{k} \leq v^2 \leq \frac{\max \mathrm{E}(R^2)}{k} = v_U^2$$

Regression Explaining Growth Rate of 87 Countries: 1960 to1996
The Proposed Reporting Style

Apparent Sign Conflicts In **Bold**; Significant Results Shaded

Models Explaining Growth Rates of Real per capita Income from 1960 to 1966: 87 Countries

Data Source: Sala-i-Martin, Doppelhofer and Miller (2004)

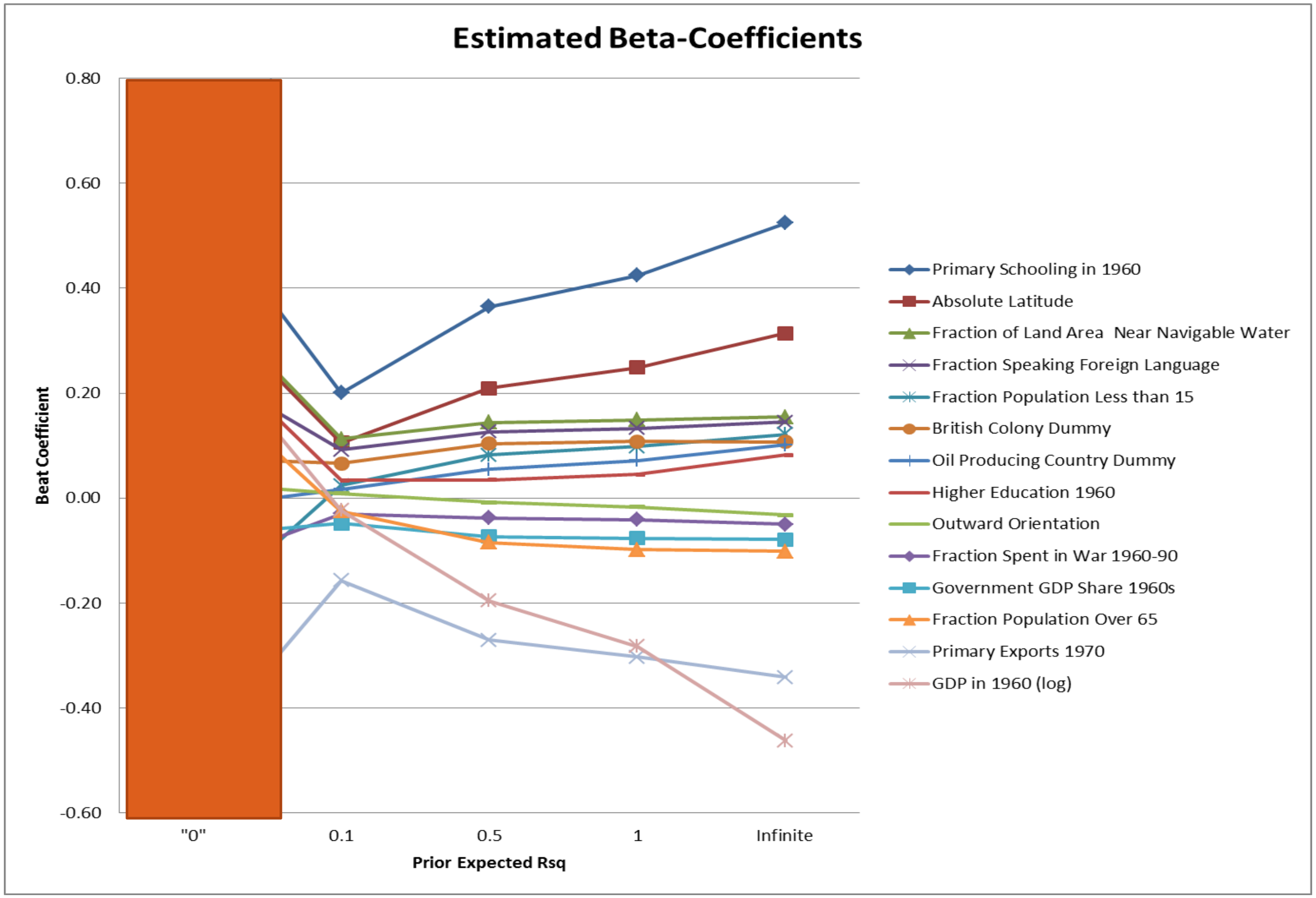
Standardized Variables (Unit variance and zero mean)

All Variables Treated the Same

Sorted First by Category and then by OLS t-value

		Prior Expected R-sq	"0"	0.1	0.5	1	Infinite	"0"	0.1	0.5	1	Infinite	(0.1, 1.0)	(0.1, 0.5)	(0.5, 1.0)
	Category	Description	b-SIMPLE	b-BAYES	b-BAYES	b-BAYES	b_OLS	t-SIMPLE	t-BAYES	t-BAYES	t-BAYES	t-OLS	s-val	s-val	s-val
1	catchup	GDP in 1960 (log)	**0.318**	-0.023	-0.195	-0.283	-0.462	3.09	-0.32	-1.60	-1.99	-2.61	-0.48	-0.47	-2.72
2	culture	Fraction Speaking Foreign Language	0.258	0.092	0.127	0.133	0.145	2.46	1.52	1.54	1.51	1.50	0.67	0.72	3.36
3	geography	Absolute Latitude	0.394	0.105	0.209	0.249	0.314	3.95	1.53	1.90	2.02	2.18	0.67	0.69	3.36
4	geography	Fraction of Land Area Near Navigable Water	0.404	0.113	0.145	0.149	0.155	4.07	1.83	1.70	1.64	1.57	0.76	0.82	3.79
5	Government	British Colony Dummy	0.076	0.066	0.104	0.108	0.107	0.70	1.11	1.30	1.27	1.17	-0.42	-0.45	-2.34
6	Government	Government GDP Share 1960s	-0.073	-0.048	-0.073	-0.076	-0.079	-0.67	-0.81	-0.94	-0.93	-0.90	-0.24	-0.25	-1.22
7	Government	Fraction Spent in War 1960-90	-0.135	-0.029	-0.038	-0.041	-0.050	-1.25	-0.50	-0.48	-0.50	-0.57	0.56	0.59	3.06
8	Government	Outward Orientation	**0.030**	0.009	-0.008	-0.017	-0.032	0.28	0.16	-0.10	-0.20	-0.36	-0.03	-0.02	-0.38
9	resources	Primary Schooling in 1960	0.574	0.201	0.365	0.424	0.525	6.47	3.03	3.61	3.81	4.14	1.34	1.40	6.89
10	resources	Primary Exports 1970	-0.491	-0.157	-0.271	-0.302	-0.342	-5.19	-2.34	-2.63	-2.65	-2.61	0.17	0.17	0.82
11	resources	Oil Producing Country Dummy	**-0.019**	0.017	0.055	0.072	0.102	-0.18	0.28	0.70	0.86	1.14	0.16	0.18	0.66
12	resources	Higher Education 1960	0.308	0.034	0.035	0.045	0.082	2.98	0.51	0.34	0.40	0.62	-0.17	-0.17	-0.83
13	resources	Fraction Population Less than 15	**-0.228**	0.024	0.082	0.099	0.122	-2.16	0.34	0.63	0.62	0.53	-0.96	-1.00	-4.81
14	resources	Fraction Population Over 65	**0.234**	-0.025	-0.085	-0.098	-0.102	2.22	-0.35	-0.65	-0.62	-0.44	0.29	0.30	1.94
		R-squared		0.210	0.378	0.436	0.537								
		Mean Squared Coefficient		0.008	0.026	0.036	0.058								

Can you see the conflict between the signs of the simple regressions and the signs of the multiple regression?



t-values and s-values are related