

Multivariate Regression

Statistics 407, ISU

Definition

Multivariate regression refers to modeling multiple responses, in addition to possibly multiple predictors.

$$\mathbf{Y}_{n \times q} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times q} + \boldsymbol{\epsilon}_{n \times q}$$

Examples might include (1) a collection of psychological variables and a collection of test scores, (2) a collection of health variables and a collection of eating habits variables, or (3) measurements on the physical characteristics of plants and variables describing the growing conditions.

Example

Relationships between properties of pulp fiber and paper, Lee (1992) unpublished

Y_1 = breaking length (BL) X_1 = arithmetic fiber length (AFL)
 Y_2 = elastic modulus (EM) X_2 = long fiber fraction (LFF)
 Y_3 = stress at failure (SF) X_3 = fine fiber fraction (FFF)
 Y_4 = burst strength (BS) X_4 = zero span tensile (ZST)

q=4, p=4

Example

	BL	EM	SF	BS	AFL	LFF	FFF	ZST
Mean	21.72	7.27	5.64	1.02	-0.02	39.03	26.68	1.07
Std dev	2.88	0.72	1.46	0.69	0.25	14.87	17.56	0.03

Corr(Y)

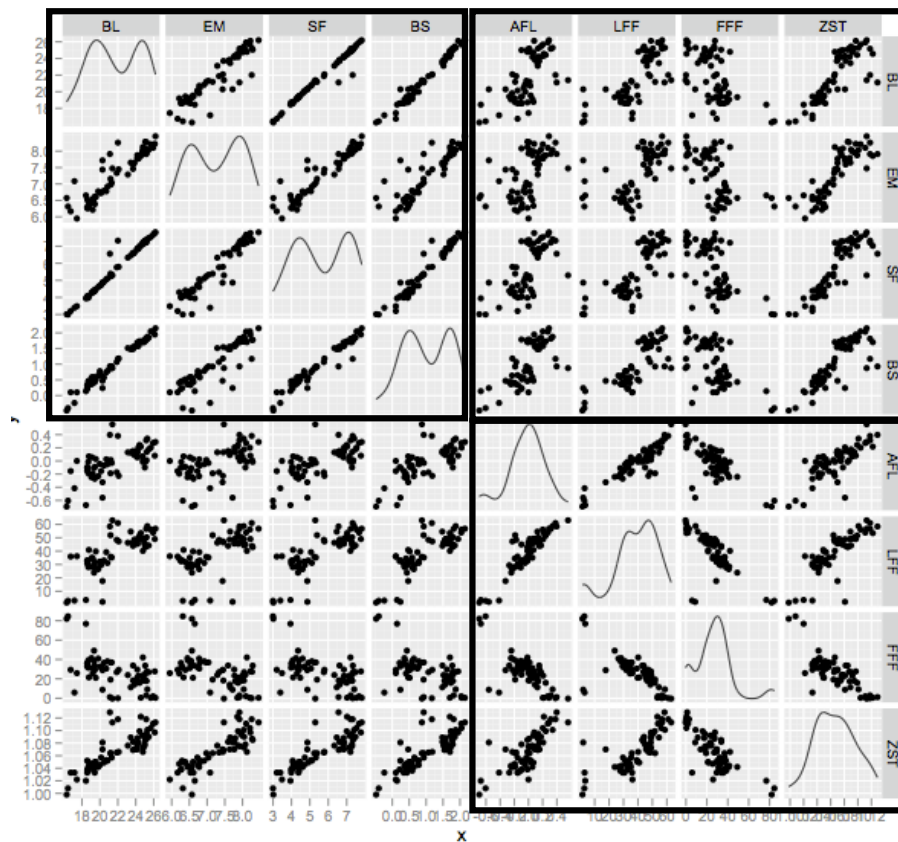
	BL	EM	SF	BS
BL	1.00	0.91	0.98	0.99
EM	0.91	1.00	0.94	0.87
SF	0.98	0.94	1.00	0.97
BS	0.99	0.87	0.97	1.00

Corr(X)

	AFL	LFF	FFF	ZST
AFL	1.00	0.91	-0.73	0.78
LFF	0.91	1.00	-0.71	0.79
FFF	-0.73	-0.71	1.00	-0.78
ZST	0.78	0.79	-0.78	1.00

Corr(Y,X)

	AFL	LFF	FFF	ZST
BL	0.65	0.74	-0.54	0.82
EM	0.54	0.61	-0.56	0.85
SF	0.68	0.76	-0.57	0.87
BS	0.71	0.80	-0.56	0.81



Response variables strongly positively correlated, outliers. Predictors somewhat correlated both negative and positive, outliers. Response vs predictors some correlation, clusters and outliers.

Fit separate models

	(Intercept)	AFL	LFF	FFF	ZST
BL	-74.23	-3.12	0.10**	0.05**	85.08***
EM	-24.01	-1.18**	0.01	0.01	28.75***
SF	-45.76	-1.49	0.05**	0.03**	45.80***
BS	-17.73	-0.55	0.03***	0.01*	16.22***

- Coefficients are similar for each response
- Significance varies, though
- ZST is most important, AFL least important.

Multivariate regression

	(Intercept)	AFL	LFF	FFF	ZST
BL	-74.23	-3.12	0.10	0.05	85.08
EM	-24.01	-1.18	0.01	0.01	28.75
SF	-45.76	-1.49	0.05	0.03	45.80
BS	-17.73	-0.55	0.03	0.01	16.22

WAIT!!! These are the same as fitting separate models.

Multivariate tests of coefficients

$$H_o : \beta = 0$$

$$H_A : \text{at least one } \beta \text{ is not } 0$$

The matrix of coefficients are jointly tested for significance. This is the difference from fitting separate models.

Multivariate tests of coefficients

$$\mathbf{E} = (\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}})' \quad \mathbf{H} = \hat{\mathbf{Y}}\hat{\mathbf{Y}}'$$

$$\text{Wilks } \Lambda = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|}$$

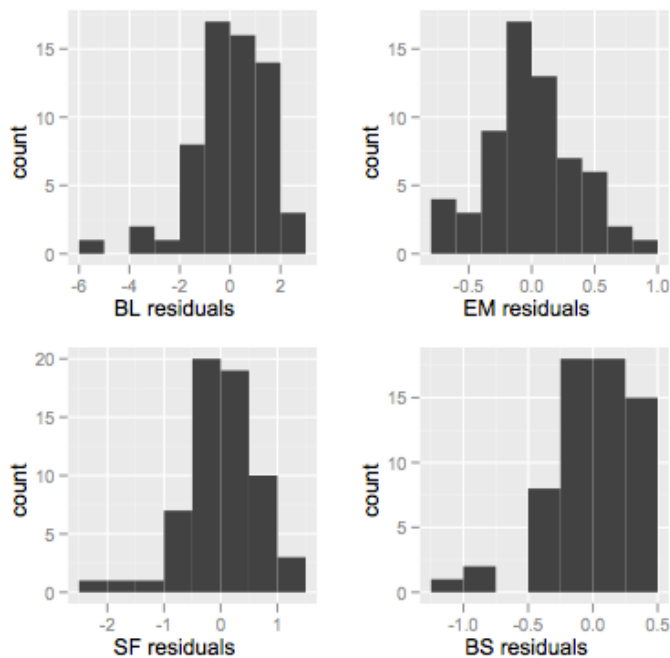
Yes, its the same Wilks. Here the comparison is between the variation due to the model (model sum of squares), and the variation due to error (residual sum of squares).

Example

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
(Intercept)	1	0.00	54983.25	4	54	0.0000
AFL	1	0.21	51.51	4	54	0.0000
LFF	1	0.54	11.49	4	54	0.0000
FFF	1	0.69	6.10	4	54	0.0004
ZST	1	0.30	30.80	4	54	0.0000
Residuals	57					

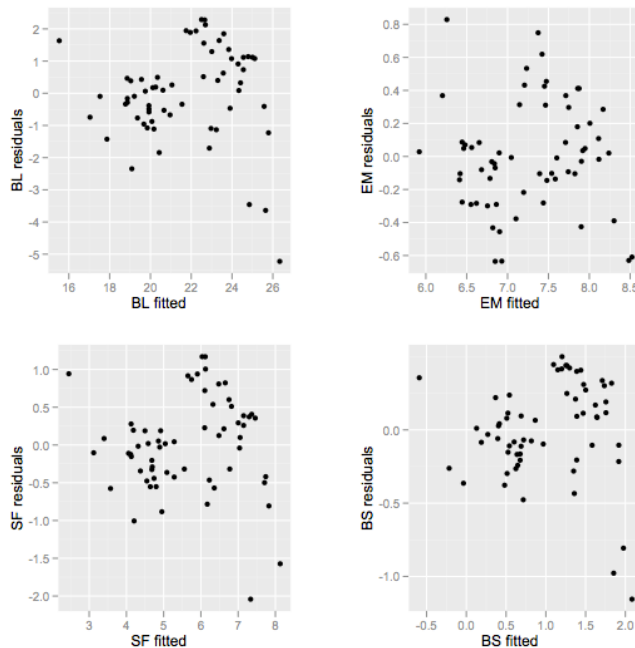
- All four predictors are very important for modeling the paper quality.
- Interestingly AFL is most important, followed by ZST, and FFF is least important

Diagnostics



Residuals on all variables should be samples from a normal distribution. Some problems with outliers here.

Diagnostics



Residuals vs fitted. Indicates some problems with a few outliers.

Alternatives

- PCA on the response variables, use a linear combination of these variables as the response. (And similarly PCA on predictors.)
- Canonical correlation analysis - which combination of responses is most correlated with which combination of predictors.
- Useful explanation: <http://www.ats.ucla.edu/stat/Stata/dae/mvreg.htm>

This work is licensed under the Creative Commons Attribution-Noncommercial 3.0 United States License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/3.0/us/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.