Multivariate Analysis Part II

Nancy Flournoy

University of Missouri

September 11, 2019
4th International Summer School on Data Science
Split, Croatia

Outline

1. Correspondence Analysis
2. Canonical Correlation
3. Sliced Regression
4. Sampling From Big Data

Correspondence Analysis Relation to PCA (Wikipedia)

Correspondence analysis (CA) was developed by Jean-Paul Benzcri. It is conceptually similar to PCA, but scales the data (which should be non-negative) so that rows and columns are treated equivalently. It is traditionally applied to contingency tables. CA decomposes the chi-squared statistic associated to this table into orthogonal factors. Because CA is a descriptive technique, it can be applied to tables for which the chi-squared statistic is appropriate or not. Several variants of CA are available including detrended correspondence analysis and canonical correspondence analysis. One special extension is multiple correspondence analysis, which may be seen as the counterpart of principal component analysis for categorical data.

Correspondence Analysis

- A technique similar to Factor Analysis for categorical data.
- To explore the structure of variables
- To reduce dimensionality
- Commonly used with two-way or multi-way contingency tables
- Similar techniques developed independently are known as optimal scaling, reciprocal averaging, optimal scoring, quantification method and homogeneity analysis.
- Details in classic text by Greenacre (1984)

CA for a Two-Way Table (www.statsoft.com)

	Smoking Category				
Staff Group	(1) None	(2) Light	(3) Medium	(4) Heavy	Row Totals
(1) Senior Managers	4	2	3	2	11
(2) Junior Managers	4	3	7	4	18
(3) Senior Employees	25	10	12	4	51
(4) Junior Employees	18	24	33	13	88
(5) Secretaries	10	6	7	2	25
Column Totals	61	45	62	25	193

Standardize frequencies to total one (divide by 193) to show how one unit of mass is distributed across the cells.

1. Think of the 4 column values in each row as coordinates in 4-D space.
2. Distances between the 4-D points summarize information about the similarities in the rows (types of employees).
3. Reduce dimensionality - optional
4. Plot results

CA Terminology

- Row and column totals of the matrix of relative frequencies are called the **row mass** and **column mass**
- Moment of inertia is the integral of mass times the squared distance to the centroid. Analogously, **inertia** is defined as the total Pearson Chi-square divided by the total sum (193 in example)
- $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$ used as a descriptive statistic.
  χ^2 distribution assumption not needed.
  f_o = observed cell frequency
  f_e = expected cell frequency if NO relationship exists between variables. $f_e = \frac{\text{row total} \times \text{column total}}{\text{table total}} =$
  $\text{table total} \times \left[\frac{\text{row total}}{\text{table total}}\right] \times \left[\frac{\text{column total}}{\text{table total}}\right]$

Correspondence Example

Eigenvalues and Inertia for all Dimensions
Input Table (Rows x Columns): 5 x 4
Total Inertia = .08519 Chi^2 = 16.442

No. of Dims	Singular Values	Eigen-Values	Perc. of Inertia	Cumulatv Percent	Chi Squares
1	.273421	.074759	87.75587	87.7559	14.42851
2	.100086	.010017	11.75865	99.5145	1.93332
3	.020337	.000414	.48547	100.0000	.07982

1. Think

Maximum Number of Dimensions.

Since the sums of the frequencies across the columns must be equal to the row totals, and the sums across the rows equal to the column totals, there are only (no. of columns-1) independent entries in each row, and (no. of rows-1) independent entries in each column of the table (once you know what these entries are, you can fill in the rest based on your knowledge of the column and row marginal totals).

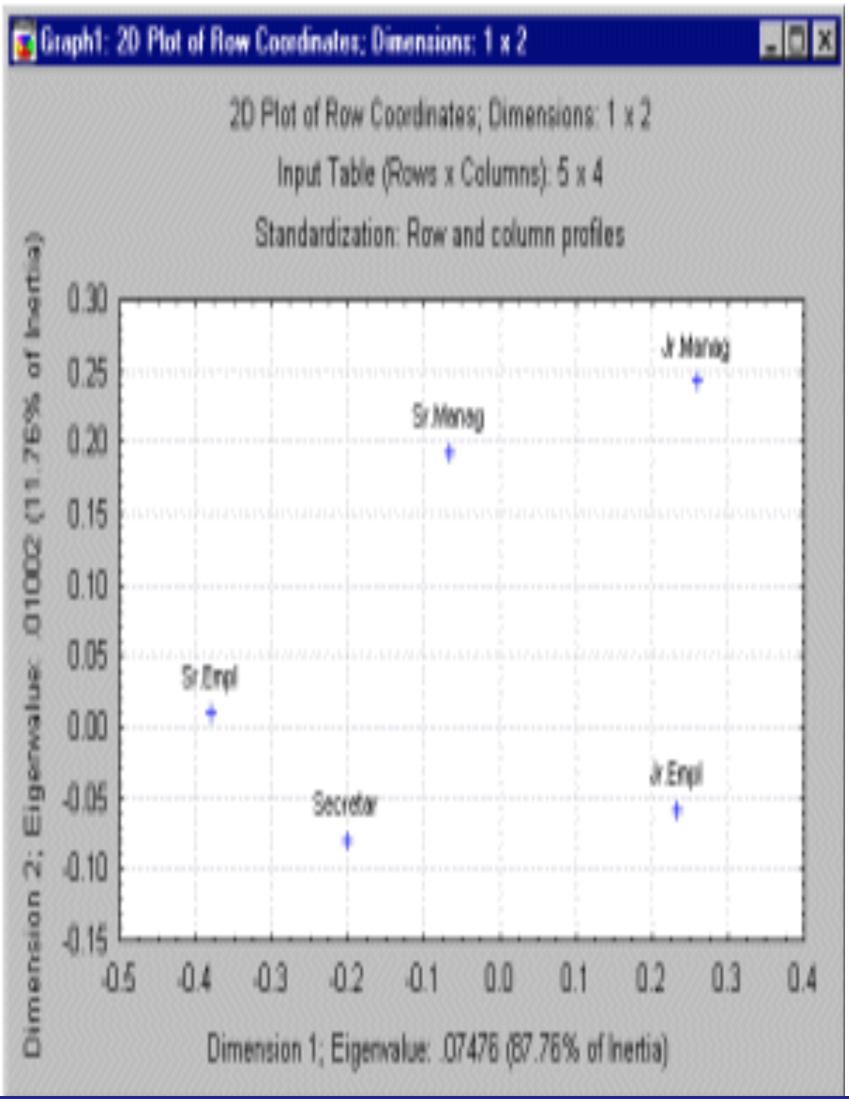
Thus, the maximum number of eigenvalues that can be extracted from a two-way table is equal to **the minimum of the number of columns minus 1, and the number of rows minus 1**.

If you extract (i.e., interpret) the maximum number of dimensions that can be extracted, then you can reproduce exactly all information contained in the table.

PC1 and PC2 Loadings from Correspondence Example

(1) Senior Managers	-.065768	.193737
(2) Junior Managers	.258958	.243305
(3) Senior Employees	-.380595	.010660
(4) Junior Employees	.232952	-.057744
(5) Secretaries	-.201089	-.078911

PC1: Senior employees and Secretaries are relatively close together.



Interpreting Factor Loadings

PC1: High percentages of Senior employees, Secretaries and Senior Managers do not smoke.
Managers and Employees differ in PC2.

Input Table Construction

One could also start with a two-way table in which the columns or rows sum to one.

CA Standardizing Row Totals

Percentages of Row Totals					
	Smoking Category				
Staff Group	(1) None	(2) Light	(3) Medium	(4) Heavy	Row Totals
(1) Senior Managers	36.36	18.18	27.27	18.18	100.00
(2) Junior Managers	22.22	16.67	38.89	22.22	100.00
(3) Senior Employees	49.02	19.61	23.53	7.84	100.00
(4) Junior Employees	20.45	27.27	37.50	14.77	100.00
(5) Secretaries	40.00	24.00	28.00	8.00	100.00

Smoking category	Dim. 1	Dim. 2
None	-.393308	.030492
Light	.099456	-.141064
Medium	.196321	-.007359
Heavy	.293776	.197766

Canonical Correlation Analysis

Harold Hotellng, 1936

Goal: Assess the relationship between two sets of variables:
If $\mathbf{X} = (X_1, \ldots X_r)^T$ and $\mathbf{Y} = (Y_1, \ldots Y_s)^T$, canonical correlation analysis finds linear combinations of $\mathbf{X}$ and $\mathbf{Y}$ that have maximum correlation with each other.

Canonical Correlation Analysis (www.statsoft.com)

Example 1. A researcher has collected data on three psychological variables, four academic variables (standardized test scores) and gender for 600 college freshman.

She is interested in **how the set of psychological variables relates to the academic variables and gender.**

In particular, the researcher is interested in how many dimensions (canonical variables) are necessary to understand the association between the two sets of variables.

Canonical Correlation Analysis

Example 2. A researcher is interested in exploring associations among factors from two multidimensional personality tests, the MMPI and the NEO.

She is interested in **what dimensions are common between the tests and how much variance is shared** between them.

She is specifically interested in finding whether the neuroticism dimension from the NEO can account for a substantial amount of shared variance between the two tests.

CCA Definition

1. Seek vectors $\mathbf{a}\,(\mathbf{a} \in \Re^r)$ and $\mathbf{b}\,(\mathbf{b} \in \Re^s)$ such that the random variables $\mathbf{a}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{Y}$ maximize the correlation

$$\rho = \text{corr}(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}).$$

   This gives the **first pair of canonical variables**.
2. Then find vectors maximizing the same correlation subject to the constraint that they are uncorrelated with the first pair of canonical variables.
   This gives the **second pair of canonical variables**.
3. This procedure may be continued up to $\min\{r, s\}$ times.

CCA Computation

To get the first Cannonical Correlation, the problem is to maximize

$$\rho = \frac{\mathbf{a}^T \Sigma_{XY} \mathbf{b}}{\sqrt{\mathbf{a}^T \Sigma_{XX} \mathbf{a}} \sqrt{\mathbf{b}^T \Sigma_{YY} \mathbf{b}}}.$$

There are two approaches:

- Eigenvalue Analysis
- Using SVD on a correlation matrix

CCA by Eigenvalue Analysis

Change basis: $\mathbf{c} = \Sigma_{XX}^{1/2}\mathbf{a}$; $\mathbf{d} = \Sigma_{XX}^{1/2}\mathbf{b}$ and use Cauchy-Schwarz inequality to bound ρ:

$$\rho \leq \frac{(\mathbf{c}^T \Sigma_{\boldsymbol{XX}}^{-1/2} \boldsymbol{\Sigma_{XY}} \Sigma_{\boldsymbol{YY}}^{-1} \Sigma_{\boldsymbol{YX}} \Sigma_{\boldsymbol{XX}}^{-1/2} \mathbf{c})^{1/2}}{(\mathbf{c}^T \mathbf{c})^{1/2}}.$$

The upper bound is attained

1. If the vectors $\mathbf{d}$ and $\Sigma_{\boldsymbol{YY}}^{-1/2} \Sigma_{\boldsymbol{YX}} \Sigma_{\boldsymbol{XX}}$ are collinear, which is,
2. If $\mathbf{c}$ is the eigenvector with the maximum eigenvalue for the matrix $\Sigma_{\boldsymbol{XX}}^{-1/2} \boldsymbol{\Sigma_{XY}} \Sigma_{\boldsymbol{YY}}^{-1} \Sigma_{\boldsymbol{YX}} \Sigma_{\boldsymbol{XX}}^{-1/2}$.

Subsequent pairs of vectors are found using eigenvalues of decreasing magnitudes.
Orthogonality is guaranteed by the symmetry of the correlation matrices.

CCA Computation Using SVD on a correlation matrix

1. Take $\mathbf{c}$ and $\mathbf{d}$ to be the left and right singular vectors of the correlation matrix of $\mathbf{X}$ and $\mathbf{Y}$ corresponding to the highest singular value in SVD.

2. Find subsequent pairs corresponding to decreasing singular values.

The square root of the eigenvalues can be interpreted as correlation coefficients. They are called *canonical correlations*.

CCA: Selecting Vector Pairs

Mendoza, Markos, and Gonter (1978) describing testing correlations for significance and only retaining pairs with correlations that are statistically significant for subsequent interpretation. They show

- the testing procedure will detect strong canonical correlations most of the time, even with samples of relatively small size (e.g., n = 50).
- Weaker canonical correlations (e.g., R = .3) require larger sample sizes ($n > 200$) to be detected at least 50% of the time.

Note that canonical correlations of small magnitude are often of little practical value, as they account for very little actual variability in the data.

CCA: Interpreting Canonical Weights

- Look at the simple correlations between transformed variables. These are called *canonical factor loadings*.
- Look at the weights for each variable in a pair of transformations.These are called *canonical weights*.

  Interpret them as you would regression or factor analysis weights.

CCA: Extracting Factors

Suppose a satisfaction survey has two items:
(1) 'Are you satisfied with your supervisors?' and
(2) 'Are you satisfied with your bosses?"
The simple correlations between the respective sum scores with the two items may both be substantial, but they are obviously very redundant.
Weights for the weighted sums (canonical variates) in each set are computed to correlate maximally, so only one is needed. The second item will receive a negligibly small weight.

Sliced Inverse Regression (SIR)

SIR uses the inverse regression curve to perform a weighted principal component analysis and thereby identify the effective dimension reducing directions.

Sliced Inverse Regression (SIR)

Model: Given a response variable $\mathbf{Y}$ and a random vector $\mathbf{X} \in \Re^p$ of explanatory variables, $\mathbf{Y} = f(\beta_1^T\mathbf{X}, \ldots, \beta_k^T\mathbf{X}, \epsilon)$, where $\beta_1, \ldots, \beta_k$ are unknown projection vectors, k the an unknown dimension ($k \leq p$). f is unknown on $\Re^{k+1}$.

The error ϵ has $E[\epsilon|\mathbf{X}] = 0$ and finite variance σ^2.

$\Rightarrow$The conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ depends on $\mathbf{X}$ only through the k dimensional random vector $(\beta_1, \ldots, \beta_k)$.

SIR sketch

1. $m(y) = E[\mathbf{X}|\mathbf{Y} = y] = E[X]$ is the centered inverse regression, a curve in $\Re^p$ consisting of p $1 - D$ regressions.
2. Estimate $m(y)$ by dividing the range of $\mathbf{Y}$ into H non-overlapping intervals (**slices**).
3. $\hat{m}_h$ is sample mean of slice h.
4. Conduct PCA on $\hat{m}_h$s after standardizing $\mathbf{X}$ to $\mathbf{Z} = \mathbf{\Sigma}_{xx}^{-1/2}[\mathbf{X} - E(\mathbf{X})]$.
5. See references in Wikipedia for details.

Sampling From Big Data

Big Data in Linear Regression

- In big data setting, the sample size n and dimension p can both be very large. We now focus on the case that $n \gg p$.
- For example, n may be on the order of a billion and p may be over a thousand (Raskutti and Mahoney, 2016).
- Even with a unique closed-form solution, the computational cost is too high with a computing complexity of $O(np^2)$.

Subsampling-based methods and their limitations

- Existing subsampling-based methods are often call **leveraging methods**.
- The key is to use nonuniform sampling probabilities so that influential data points are sampled with high probabilities.
- Most of the existing methods use the normalized leverage scores as subsampling probabilities.
- Calculating exact leverage scores requires $O(np^2)$ time; Calculating approximate leverage scores requires $O(np \log n/\epsilon^2)$, where $\epsilon \in (0, 0.5]$ (Drineas et al. 2012).
- Another limitation is that information obtained is typically at the scale of the subdata size and not the full data size.

References on subsampling-based methods

- Drineas, P., Mahoney, M. W., and Muthukrishnan, S. (2006). Sampling algorithms for L_2 regression and applications. In **Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms**, 1127–1136.
- Rokhlin, V. and Tygert, M. (2008) A fast randomized algorithm for overdetermined linear least-squares regression. **Proceedings of the National Academy of Sciences of the United States of America**, 105(36):13212–13217.
- Drineas, P., Mahoney, M., Muthukrishnan, S., and Sarlos, T. (2011). Faster least squares approximation. **Numerische Mathematik 117**, 219–249.
- Dhillon, P., Lu, Y., Foster, D. P. and Ungar, L. (2013) New subsampling algorithms for fast least squares regression. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, **Advances in Neural Information Processing Systems 26**, pages 360–368.
- Ma, P., Mahoney, M., and Yu, B. (2015). A statistical perspective on algorithmic leveraging. **Journal of Machine Learning Research 16**, 861–911.
- Yang, T., Zhang, L., Jin, R., and Zhu, S. (2015). An explicit sampling dependent spectral error bound for column subset selection. In **Proceedings of The 32nd International Conference on Machine Learning**, 135-143.
- Raskutti, G. and Mahoney, M. (2016). A Statistical Perspective on Randomized Sketching for Ordinary Least-Squares. **Journal of Machine Learning Research 17**, 1–31.

Design Theory Based Sampling

IBOSS:

Wang, HaiYing, Min Yang, and John Stufken.
Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association* 114.525 (2019): 393-405.

What difference does design make?

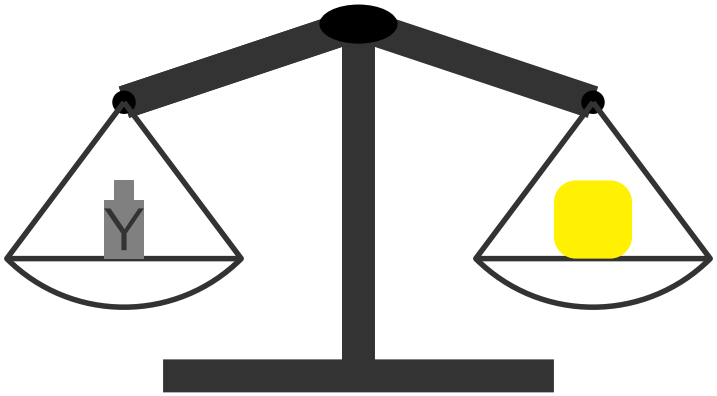
Does it Matter How You Sample?

- Four gold bars have weights A, B, C and D.
- \$1 charge each time the scale is used.
- The scale may not be very accurate: error$\sim N(0, \sigma^2)$.

How would you weight the gold bars if you have only \$4?

Method 1: Weigh each gold bar individually

Observe weights Y_1, Y_2, Y_3, Y_4.

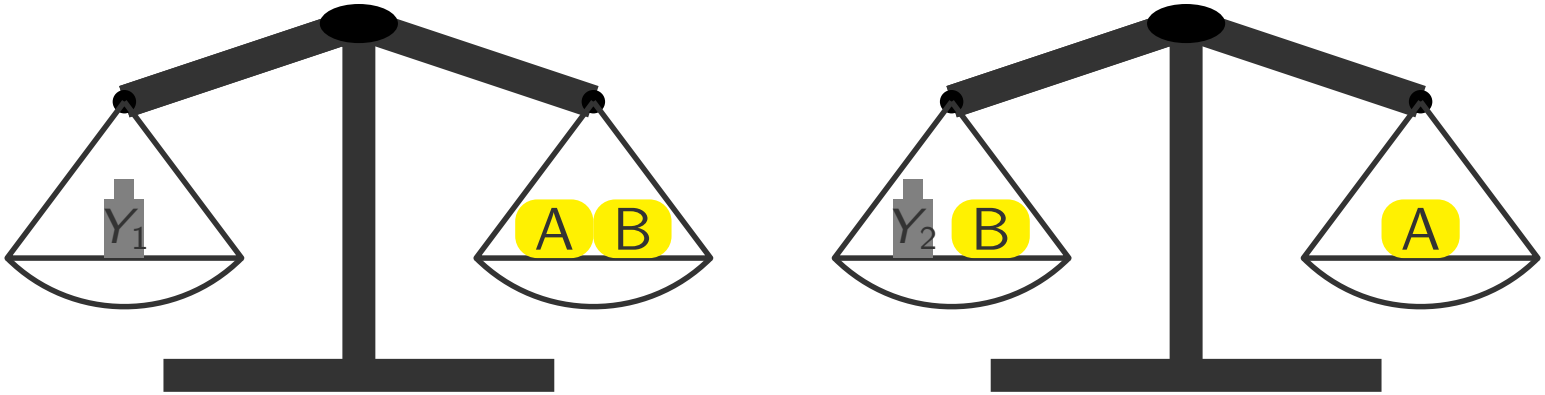


$$Y_1 = A + \varepsilon_1; \quad Y_2 = B + \varepsilon_2; \quad \widehat{A} = Y_1 \quad \widehat{B} = Y_2;$$
$$Y_3 = C + \varepsilon_3; \quad Y_4 = D + \varepsilon_4. \quad \widehat{C} = Y_1 \quad \widehat{D} = Y_4.$$

How precise is this method? $Var(Y_i) = Var(\epsilon_i) = \sigma^2$.

- $\text{Var}(\widehat{A})=\text{Var}(\widehat{B})=\text{Var}(\widehat{C})=\text{Var}(\widehat{D})=\sigma^2$

Method 2: Weigh $A+B$, $A-B$, $C+D$, $C-D$



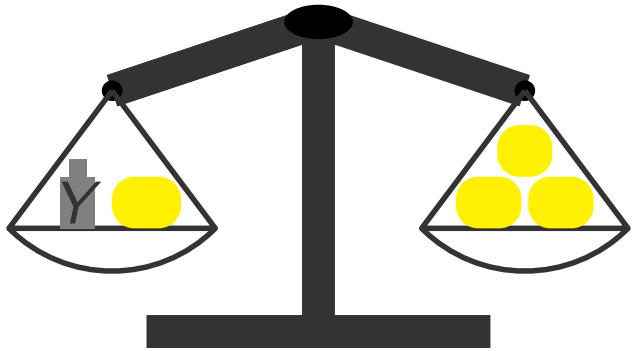
$$Y_1 = A + B + \varepsilon_1; \quad Y_2 = A - B + \varepsilon_2;$$
$$Y_3 = C + D + \varepsilon_3; \quad Y_4 = C - D + \varepsilon_4.$$

$$\widehat{A} = \frac{Y_1 + Y_2}{2};$$
$$\widehat{B} = \frac{Y_1 - Y_2}{2}.$$

How precise is this method?

$$Var(\widehat{A}) = Var(\widehat{B}) = \frac{Var(Y_1) + Var(Y_2)}{4} = \frac{\sigma^2}{2}; \quad Var(\widehat{C}) = Var(\widehat{D}) = \frac{\sigma^2}{2}$$

Method 3: Weigh the difference between one gold bar and the sum of the other three



$$Y_1 = A + B + C - D + \varepsilon_1; \qquad Y_2 = A + B + D - C + \varepsilon_2$$
$$Y_3 = A + C + D - B + \varepsilon_3; \qquad Y_4 = B + C + D - A + \varepsilon_4$$

How precise is this method?

- $\text{Var}(\widehat{A}) = \text{Var}(\widehat{B}) = \text{Var}(\widehat{C}) = \text{Var}(\widehat{D}) = \frac{1}{4}\sigma^2$

Comparison

The three methods corresponded to three designs.

- All methods give unbiased estimates of the weights.
- Method 3 has smallest variances $\frac{1}{4}\sigma^2$.
- To achieve the same precision
  - Method 1 needs \$16
  - Method 2 needs \$8
  - Method 3 needs \$4

Are there any designs better than Method 3?

No. It can be shown that Method 3 is the optimal design.

Typical Design Goals:

Sample/Treat so as to

- Reduce sample size, cost and/or subject risk needed to achieve a specific precision for a desired estimator or set of estimators.
- Increase precision for given cost and/or sample size and/or limit to subject risk.

Some Optimal Design Criterion

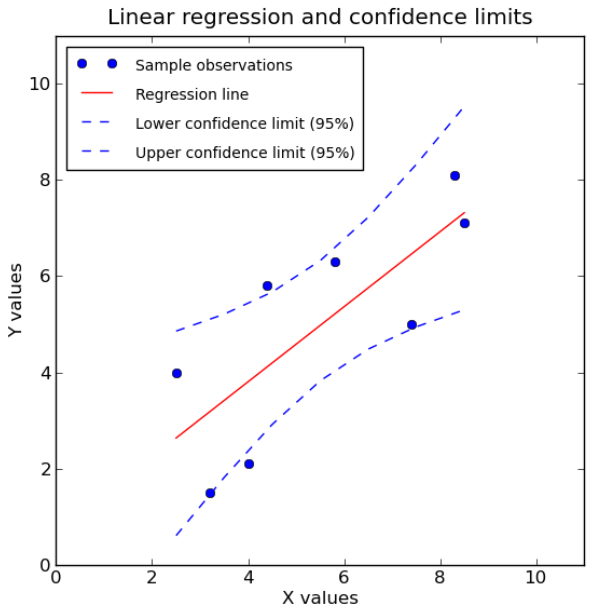
Optimality criterion: specify how to collect samples, for example, so as to

- minimize variance of estimator
- minimize variance of estimator given constraint on cost or subject risk
- minimize confidence ellipsoid for multiple estimators
- minimize maximum eigenvalue of information matrix

IBOSS uses D-optimal Sampling Criterion

Sample subjects (items, whatever) in such a way as to maximize the determinant of the multivariate confidence ellipsoid.

Minimize Confidence Ellipsoid in 2-D Linear Model



Confidence Ellipsoid is minimized if equal amounts of data are sampled at the extreme values of the permissible x-values.

In this example,
sample y-values for which
$x = 2$ and $x = 8$