ORIGIN ≡ 0

Prepared by:
Wm Stein

Principal Axis Regression

Linear Regression in the "usual" or "Type I" mode, as described by Sokal & Rohlf (*Biometry* 3rd Edition 1995) involves specifying in advance which of two variables is to be considered fixed and independent (X) versus dependent with natural variation (Y). This prior assignment is useful for prediction of Y from X, and for calculating confidence intervals, since all variation from a trend line is interpreted to occur in Y alone with a single Normally distributed error term ε. For futher details, I highly recommend reading Sokal & Rohlf's discussion of this in their chapters on Regression and Correlation. Type II regressions, by constrast to Type I, are useful primarily for finding "best fit" trend lines between variables each of which are considered to have natural variation. Rather than prediction of Y from X, the main object is to describe a relationship between variables in a bivariate sense.

Principal Axis (PA) Regression (also called Major Axis (MA) Regression) is derived from principal axes (as in PCA) of variables using either standardized (correlation matrix) or unstandardized (covariance matrix) data. This choice, as with all PCA methods, centers around whether the researcher believes the scale different variables are measured interferes or enhances interpretation of variable relationships. It is typical for the covariance matrix to be used when "slopes" have potential meaning as, for instance, in allometry. Otherwise, use of the correlation matrix is considered "conservative".

Assumptions:

- **Type II regressions are symmetrical with regard to natural variation.**
  Each variable X & Y is considered to be derived from a population of possible values:
- **$Y_1, Y_2, Y_3, \ldots, Y_n$ is a random sample ~ $N(\mu_Y, \sigma_Y^2)$.**
- **$X_1, X_2, X_3, \ldots, X_n$ is a random sample ~ $N(\mu_X, \sigma_X^2)$**
- **Values of X_i are matched to Y_i**

Model:

$$Y = \alpha + \beta X$$

here: α is the y intercept of the regression line (translation)
β is the slope of the regression line (scaling coefficient)
where: each variable has its own error term. This will not be analyzed explicitly since no predictions are formed.

Example: **Sokal & Rohlf p. 546 Box 14.12**

K := READPRN("c:/RData/SR15.2.txt")

$X := K^{\langle 0 \rangle}$ $Y := K^{\langle 1 \rangle}$ **< assigning variables X & Y from matrix K**

$X_{bar} := \text{mean}(X)$ $X_{bar} = 195.58333$ **< mean for X = first column in K**

$Y_{bar} := \text{mean}(Y)$ $Y_{bar} = 12.0475$ **< mean for Y = second column in K**

$n := \text{length}\left(K^{\langle 0 \rangle}\right)$ $n = 12$ **< number of paired observations**

K =

	0	1
0	159	14.4
1	179	15.2
2	100	11.3
3	45	2.5
4	384	22.7
5	230	14.9
6	100	1.41
7	320	15.81
8	80	4.19
9	220	15.39
10	320	17.25
11	210	9.52

Least Squares Estimation of the Regression Line:

Sums of Squares and Cross Products corrected for mean location:

$i := 0 .. n - 1$

$$L_{xx} := \sum_i \left(X_i - X_{bar}\right)^2 \qquad L_{xx} = 1.2437 \times 10^5$$ **< corrected Sum of squares of X**

$$L_{yy} := \sum_i \left(Y_i - Y_{bar}\right)^2 \qquad L_{yy} = 462.4782$$ **< corrected Sum of squares of Y**

$$L_{xy} := \sum_i \left(X_i - X_{bar}\right) \cdot \left(Y_i - Y_{bar}\right) \qquad L_{xy} = 6561.6175$$ **< corrected Sum of cross products**

Type I "Simple" Regression of Y on X:

Estimated Regression Coefficients for Y = $\alpha + \beta X$:

$b := \frac{L_{xy}}{L_{xx}}$ $b = 0.05276$ **< sample estimate of β**

$a := Y_{bar} - b \cdot X_{bar}$ $a = 1.72866$ **< sample estimate of α**

Estimated Values of Y:

$Y_{hat_i} := a + b \cdot X_i$

Residuals:

$\varepsilon_i := Y_{hat_i} - Y_i$

Mean Square for Error:

$$MSE := \frac{\sum_i (\varepsilon_i)^2}{n-2}$$ **< from ANOVA for Regression Standard Table**

$Y_{hat} =$

	0
0	10.1174
1	11.1726
2	7.0046
3	4.1028
4	21.9882
5	13.8633
6	7.0046
7	18.6116
8	5.9494
9	13.3357
10	18.6116
11	12.8081

$\varepsilon =$

	0
0	-4.2826
1	-4.0274
2	-4.2954
3	1.6028
4	-0.7118
5	-1.0367
6	5.5946
7	2.8016
8	1.7594
9	-2.0543
10	1.3616
11	3.2881

Standard Error of slope:

$s_b := \sqrt{\frac{MSE}{L_{xx}}}$ $s_b = 0.00967$ **< See SR p. 467 & 471 for calculation of s_b**

Type II RMA - Reduced Major Axis Regression:

Estimated Regression Coefficients:

$b_\nu := \sqrt{\frac{L_{yy}}{L_{xx}}}$ $b_\nu = 0.06098$ **< RMA regression estimate of slope**

$a_\nu := Y_{bar} - b_\nu \cdot X_{bar}$ $a_\nu = 0.12077$ **< RMA regression intercept. Note that it differs slightly from that given in SR.**

Estimated Values of Y:

$Y\nu_{hat_i} := a_\nu + b_\nu \cdot X_i$

Residuals:

$\varepsilon_{\nu_i} := Y\nu_{hat_i} - Y_i$

Mean Square for Error:

$$MSE_\nu := \frac{\sum_i (\varepsilon_{\nu_i})^2}{n-2}$$

$Y\nu_{hat} =$

	0
0	9.8166
1	11.0362
2	6.2188
3	2.8649
4	23.5372
5	14.1462
6	6.2188
7	19.6345
8	4.9992
9	13.5364
10	19.6345
11	12.9266

$\varepsilon_\nu =$

	0
0	-4.5834
1	-4.1638
2	-5.0812
3	0.3649
4	0.8372
5	-0.7538
6	4.8088
7	3.8245
8	0.8092
9	-1.8536
10	2.3845
11	3.4066

Standard Error of slope:

$s_b := \sqrt{\frac{MSE_\nu}{L_{xx}}}$ $s_b = 0.01001$ **< Note that my calculations use residuals in Y calculated from the RMA predicted values, rather than the Type I predicted values. This approach differs from that of SR where values are simply taken from Type I regression above. I'm not sure the value of this.**

Multivariate items needed for PCA:

$X := K^T$

$n := \text{cols}(X)$ $n = 12$ $p := \text{rows}(X)$ $p = 2$

$i := 0..n-1$ $j := 0..p-1$ $k := 0..p-1$

$1_i := 1$ $I := \text{identity}(n)$

Mean Vector:

$X_{bar} := \frac{1}{n}\cdot X\cdot 1$ $X_{bar} = \begin{pmatrix} 195.58333 \\ 12.0475 \end{pmatrix}$ **< Mean Vector**

Covariance Matrix:

$S := \frac{1}{n-1}\cdot X\cdot\left(I - \frac{1}{n}\cdot 1\cdot 1^T\right)\cdot X^T$ $S = \begin{pmatrix} 11306.26515 & 596.51068 \\ 596.51068 & 42.04348 \end{pmatrix}$ **< Covariance Matrix**

Correlation Matrix:

$D_{j,k} := \begin{cases} \frac{1}{\sqrt{S_{j,k}}} & \text{if } j = k \\ 0 & \text{otherwise} \end{cases}$ $D = \begin{pmatrix} 0.0094 & 0 \\ 0 & 0.15422 \end{pmatrix}$ **< Inverse Standard Deviation Matrix**

$R := D\cdot S\cdot D$ $R = \begin{pmatrix} 1 & 0.86519 \\ 0.86519 & 1 \end{pmatrix}$ **< Correlation Matrix**

Eigenanalysis for Standardized Data:

$\Lambda_s := \text{reverse}(\text{sort}(\text{eigenvals}(R)))$ $\Lambda_s = \begin{pmatrix} 1.86519 \\ 0.13481 \end{pmatrix}$ **< Eigenvalues in numerical order**

$E_s^{\langle j\rangle} := \text{eigenvec}\left(R, \Lambda_{s_j}\right)$ $E_s = \begin{pmatrix} 0.70711 & -0.70711 \\ 0.70711 & 0.70711 \end{pmatrix}$ **< Eigenvectors in corresponding columns. Directions indicate 45 degrees from the original axes. This makes sense because the variables have equal variance =1 when standardized.**

$\sqrt{\frac{S_{1,1}}{S_{0,0}}} = 0.06098$ **< Slope coefficient calculated directly from covariance matrix S.**

Eigenanalysis for Unstandardized Data:

$\Lambda := \text{reverse}(\text{sort}(\text{eigenvals}(S)))$ $\Lambda = \begin{pmatrix} 11337.76601 \\ 10.54261 \end{pmatrix}$ **< Eigenvalues in numerical order**

$E^{\langle j\rangle} := \text{eigenvec}\left(S, \Lambda_j\right)$ $E = \begin{pmatrix} 0.9986085 & -0.0527351 \\ 0.0527351 & 0.9986085 \end{pmatrix}$ **< Eigenvectors in corresponding columns**

Estimated Regression Coefficients:

$B_j := \frac{S_{0,1}}{\Lambda_j - S_{0,0}}$ $B = \begin{pmatrix} 18.93633 \\ -0.05281 \end{pmatrix}$ **< Slopes calculated as in Box on SR p. 589 but note that these are just the ratio of components of the Eigenvectors - i.e., "slopes".**

$\frac{E_{0,0}}{E_{1,0}} = 18.93633$ $\frac{E_{0,1}}{E_{1,1}} = -0.05281$ **< eigenvector ratios**

$Int := X_{bar_0} - B\cdot X_{bar_1}$ $Int = \begin{pmatrix} -32.5521 \\ 196.2195 \end{pmatrix}$ **< Intercepts**

Confidence Interval for Slope:

$n = 12$

$\Lambda = \begin{pmatrix} 11337.76601322 \\ 10.5426133 \end{pmatrix}$ $E = \begin{pmatrix} 0.99860854 & -0.05273506 \\ 0.05273506 & 0.99860854 \end{pmatrix}$

$\alpha := 0.05$ **< Set value as desired**

$qF(1-\alpha, 1, n-2) = 4.9646$

$$H := \frac{qF(1-\alpha,1,n-2)}{\left[\left(\frac{\Lambda_0}{\Lambda_1}+\frac{\Lambda_1}{\Lambda_0}-2\right)\cdot(n-1)\right]}$$

$H = 0.0004204561$

$$A := \sqrt{\frac{H}{1-H}}$$

$A = 0.0205093382$

$$L := \begin{pmatrix} \dfrac{B_0 - A}{1 + B_0 \cdot A} \\ \dfrac{B_0 + A}{1 - B_0 \cdot A} \end{pmatrix}$$

$L = \begin{pmatrix} 13.6244648 \\ 30.9940441 \end{pmatrix}$ **< confidence for slope B_0**

Prototype in R:

COMMANDS:

> K=read.table("c:/2008Morphometrics/SR15.2.txt")
> K

	V1	V2
1	159	14.40
2	179	15.20
3	100	11.30
4	45	2.50
5	384	22.70
6	230	14.90
7	100	1.41
8	320	15.81
9	80	4.19
10	220	15.39
11	320	17.25
12	210	9.52

< returns data in K

> S=cov(K)
> S

	V1	V2
V1	11306.2652	596.51068
V2	596.5107	42.04348

^ returns covariance matrix S

> R=cor(K)
> R

	V1	V2
V1	1.0000000	0.8651857
V2	0.8651857	1.0000000

^ returns correlation matrix R

> eigen(S)

\$values
[1] 11337.76601 10.54261 **< returns eigenvalues**

\$vectors

	[,1]	[,2]
[1,]	0.99860854	-0.05273506
[2,]	0.05273506	0.99860854

< returns eigenvectors

> e\$vectors[1,1] **[1] 0.9986085**
> e\$vectors[2,1] **[1] 0.05273506**
< extracting first eigenvector

> b0=e\$vectors[1,1]/e\$vectors[2,1]
> b0 **[1] 18.93633** **< calculating first slope**

> b1=e\$vectors[1,2]/e\$vectors[2,2] **[1] -0.05280855** **< calculating second slope**
> b1

$B = \begin{pmatrix} 18.93632903 \\ -0.05280855 \end{pmatrix}$

```
> X0=K$V1
> X1=K$V2
> Int0=mean(X0)- b0*mean(X1)
> Int1=mean(X0)- b1*mean(X1)
> Int0
[1] -32.55209          < calculating intercepts
> Int1
[1] 196.2195
```

$$\mathrm{Int} = \begin{pmatrix} -32.5520906 \\ 196.2195443 \end{pmatrix}$$