Variance & Covariance

I. Measures of Central Tendency

Mode: Most frequent observation

Median: Middle of rank ordered X_i

Mean: Arithmetic = $\overline{X} = \sum_{1}^{n} (X_i)/N$

$$\text{Geometric} = \sqrt[n]{\prod_{1}^{n} (X_i)}$$

$$\text{Harmonic} = \frac{N}{\sum_{1}^{n} (1/X_i)}$$

II. Measures of Dispersion

Range: maximum - minimum

Interquartile range 75% - 25%

average absolute deviation from median

deviation score = $x = X - \overline{X}$

mean deviation = $\sum_{1}^{n} (x_i)/N =$

$$\sum_{1}^{n} (X - \overline{X})/N = \sum_{1}^{n} (X)/N - \overline{X} = 0$$

standard deviation = σ_x = root mean square deviation

variance = mean square deviation = σ^2

Standard Deviation = σ_x = root mean square deviation

$$\sigma_x = \sqrt{\sum_{1}^{n} (x_i^2)/N}$$

Variance = mean square deviation = $\sigma^2 = \sum_{1}^{n} (x_i^2)/N$

unbiased estimate of variance from a sample =

$$\sum_{1}^{n} (x_i^2)/(N-1)$$

$$V_x = \frac{\sum_{1}^{n} (X_i^2) - \left[\frac{\sum_{1}^{n} (X_i) * \sum_{1}^{n} (X_i)}{N} \right]}{(N-1)}$$

Sensitivity to transformations:

$M(X+C) = M(X) + C$
$V(X+C) = V(X)$
$V(XC) = C^2\ V(X)$

Standard Score = deviation score / standard deviation

$z = x/S_x = (X-M)/S_x$ =(a unit free index of dispersion)

$M_z = 0 \quad V_z = 1 \quad S_z = 1$

Coefficient of variation = S_x/M_x (=> ratio measurement)

Variance of Composites

$$V(X+Y) = V(x+y) = \sum_{1}^{n}((x_i+y_i)^2)/(N-1) =$$

$$\frac{\sum_{1}^{n}((x_i)^2) + \sum_{1}^{n}((y_i)^2) + 2\sum_{1}^{n}((x_i * y_i))}{(N-1)}$$

$$V(X+Y) = V_x + V_y + 2\ Cov_{xy}$$

Covariance of x and y =

$$\frac{\sum_{1}^{n}((X_i * Y_i)) - \sum_{1}^{n}((X_i)) * \sum_{1}^{n}((Y_i))}{(N-1)}$$

V(x+y) a visual representation

	x	y
x	V_x	C_{xy}
y	C_{xy}	V_y

Variance of Composites: an example
Standard deviation of GRE Verbal = 100
Standard deviation of GRE Quant = 100
Variance of Verbal = 100 * 100 = 10,000
Variance of Quant = 100 * 100 = 10,000
Covariance of GRE Q and V = 6,000

Variance of GRE (V + Q) $= V_v + V_q + 2\ C_{vq}$

	Verbal	Quantitative
Verbal	10,000	6,000
Quantitative	6,000	10,000

$V_{(V+Q)} = 32{,}000 \implies SD_{(V+Q)} = 179$

Generalization of Variance of Composites to N variables:

$V(x_1 + x_2 + \ldots x_n) =$
$Vx_1 + Vx_2 + \ldots Vx_n \quad + 2(Cx_1x_2 + Cx_1x_3 + \ldots + Cx_ix_j \ + \ldots$
)
(n terms) (n * (n-1) terms

Variance of N variables: (figural representation)

	x_1	x_2	x_3		... x_i		... x_j		... x_n
x_1	V_1	C_{12}	C_{13}		... C_{1i}				... C_{1n}
x_2	C_{21}	V_2	C_{23}		... C_{2i}				... C_{2n}
x_3	C_{31}	C_{32}	V_3		... C_{3i}				... C_{3n}
...					...				...
x_i					... V_i		... C_{ij}		...C_{in}
...					...				...
x_j	C_{j1}	C_{j2}	C_{j3}	...	... C_{ji}	...	... V_j		...C_{jn}
...									
x_n	C_{n1}	C_{n2}	C_{n3}		... C_{ni}		...C_{nj}		...V_n

A total of n variance terms on the diagonal and n * (n-1) = $n^2 - n$ covariance terms off the diagonal.

The variance of the composite of n variables = the sum of the n variances and the n * (n-1) covariances.

Correlation and Regression

The problem of predicting y from x:

Linear prediction $y=bx+c$ $Y=b(X-M_X) + M_y$

error in prediction = predicted y - observed y

problem is to minimize the squared error of prediction

minimize the error variance = $[\bullet(y_p-y_o)^2]/(N-1)$

$$V_e=V_{(bx-y)} = \bullet (bx-y)^2 /(N-1)=$$
$$\bullet(b^2x^2-2bxy+y^2) /(N-1)=$$
$$b^2\bullet x^2/(N-1)-2b\bullet xy/(N-1)+\bullet y^2/(N-1) ==>$$
$$V_e=b^2V_x-2bC_{xy}+V_y$$

V_e is minimized when the first derivative (w.r.t. b) = 0 ==>

$$\text{when } 2bV_x -2C_{xy}=0 ==>$$
$$b_{y.x}=C_{xy}/V_x$$

Similarly, the best $b_{x.y}$ is C_{xy}/V_y

The Pearson Product Moment Correlation Coefficient (PPMCC) is the geometric mean of these two slopes:

$$r_{xy} = \frac{C_{xy}}{\sqrt{V_xV_y}} = \frac{C_{xy}}{S_xS_y}$$

r_{xy} =

Error!

$$r_{xy} = \frac{C_{xy}}{\sqrt{V_x V_y}} = \frac{C_{xy}}{S_x S_y} = r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

1) If x and y are continuous variables, then r = **Pearson r**

2) if x and y are rank orders, then r = **Spearman r**

3) if x is continuous and y is dichotomous r = **point biserial**

4) if x and y are both dichotomous, then r = **phi** = $\sqrt{\frac{\text{chi square}}{N}}$

5)**Tetrachoric correlation** is an estimate of continuous (Pearson) based upon dichotomous data. This assumes bivariate normality.

6) **Biserial correlation** estimates continuous based upon one dichotomous and one continuous. It also assumes normality.

Calculating fomulae:

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

Covariance xy = $(\bullet XY - \bullet X \bullet Y/N)/(N-1)$

Variance X = $(\bullet X^2 - (\bullet X)^2/N)/(N-1)$

Variance Y = $(\bullet Y^2 - (\bullet Y)^2/N)/(N-1)$

$$\textbf{Correlation} = \frac{\textbf{Covariance}}{\sqrt{(\textbf{Variance X})(\textbf{Variance Y})}} =$$

$$\frac{(\bullet XY - \bullet X \bullet Y/N)/(N-1)}{\sqrt{[(\sum X^2 - (\sum X)^2/N)/(N-1)][(\sum Y^2 - (\sum Y)^2/N)/(N-1)]}}$$

$$r_{xy} = \frac{(\sum XY - \sum X \sum Y/N)}{\sqrt{\sum X^2 - (\sum X)^2/N}\,\sqrt{\sum Y^2 - (\sum Y)^2/N}}$$

Correlation

1) **Slope of regression** ($b_{xy} = C_{xy}/V_x$) reflects units of x and y but the correlation $\{r = C_{xy}/(S_x S_y)\}$ is unit free.

2) **Geometrically,** r = cosine (angle between test vectors)

3) **Correlation as prediction:**
Let y_p = predicted deviation score of y = predicted Y -M
$y_p = b_{xy}x$ and $b_{xy} = C_{xy}/V_x = rS_y/S_x \implies y_p/S_y = r(x/S_x) \implies$
predicted z score of y (z_{y_p}) = r_{xy} * observed z score of x (z_x)
predicted z score of x (z_{x_p}) = r_{xy} * observed z score of y (z_y)

4) Amount of **error variance** (residual or unexplained variance) in y given x and r

$V_e = \sum e^2/N = \sum(y-bx)^2/N = \sum\{y-(r*S_y*x/S_x)\}^2$

$V_y + V_y * r^2 - 2(r * S_y * C_{xy})/S_x$

(but $S_y * C_{xy}/S_x = V_y * r$)

$V_y + V_y * r^2 - 2(r^2 * V_y) = V_y(1-r^2) \implies$

$V_e = V_y(1-r^2) \iff V_{y_p} = V_y(r^2)$

Residual Variance = Original Variance * $(1-r^2)$
Variance of predicted scores = original variance * r^2

5) Correlation of x with predicted y = $C_x(bx)/(S_x * S_{y_p})$
but $C_{x(bx)} = bV_x = V_x*r*S_y/S_x$ and $S_{y_p} = rS_y$ and
therefore r(x with predicted y) = 1

	x	y	y_p	residual
Variance	V_x	V_y	$V_y(r^2)$	$V_y(1-r^2)$
Correlation with x	1	r_{xy}	1	0
Correlation with y	r_{xy}	1	r_{xy}	$\sqrt{(1-r^2)}$

Multiple Correlation

$$r_{xy} = \frac{C_{xy}}{\sqrt{V_x V_y}}$$

The problem of predicting y from x_1, x_2:

Linear prediction $y = b_1 x_1 + b_2 x_2 + c$

Just as the optimal b weights in regression are $b_{y.x} = C_{xy}/V_x$, so are the optimal b weights in multiple regression, however, they are corrected for the effect of the other variables:

In the two variable case the b weights (betas) are:

$$b_1 = \frac{C_{yx_1.x_2}}{V_{x_1.x_2}} \qquad b_2 = \frac{C_{yx_2.x_1}}{V_{x_2.x_1}}$$

$$b_1 = \frac{r_{x_1y} - r_{x_1x_2} * r_{x_2y}}{1 - r^2_{x_1x_2}} \qquad b_2 = \frac{r_{x_2y} - r_{x_1x_2} * r_{x_1y}}{1 - r^2_{x_1x_2}}$$

The amount of variance accounted for by the model is the sum of the product of the betas and the zero order correlations:

$$R^2 = \bullet beta_i * r_{x_iy}$$

Consider the following example:

extraversion with leadership $r = .56$ $r^2 = .31$

dominance with leadership $r = .42$ $r^2 = .18$

extraversion with dominance $r = .48$ $r^2 = .23$

$$beta_1 = \frac{.56 - .42 * .48}{1 - .48^2} = .47 \qquad beta_2 = \frac{.42 - .56 * .48}{1 - .48^2} = .20$$

$$R^2 = beta_1 * r_{x_1y} + beta_2 * r_{x_2y} = .47 * .56 + .20 * .42 = .35$$

Multiple Correlation as weighted linear composites

$$r_{xy} = \frac{C_{xy}}{\sqrt{V_x V_y}}$$

The problem is to find the Covariance of $(x_1\ x_2)$ with y

	x_1	x_2	y
x_1	V_1	C_{12}	C_{1y}
x_2	C_{21}	V_2	C_{2y}
y	C_{y1}	C_{y2}	V_y

Covariance $((x_1\ x_2)\ y) = C_{y1} + C_{y2}$

Variance $(x_1\ x_2) = V_1 + C_{12} + C_{21} + V_2$

Variance $(y) = V_y$

Standardized solution:

	z_1	z_2	z_y
z_1	1	r_{12}	r_{1y}
z_2	r_{21}	1	r_{2y}
z_y	r_{y1}	r_{y2}	1

Covariance $((x_1\ x_2)\ y) = r_{y1} + r_{y2}$

Variance $(x_1\ x_2) = 1 + r_{12} + r_{21} + 1$

Variance $(y) = 1$

Multiple Correlation is optimally weighted composite:

$$r_{xy} = \frac{C_{xy}}{\sqrt{V_x V_y}}$$

	b_1z_1	b_2z_2	z_y
b_1z_1	b_1^2	$b_1b_2r_{12}$	b_1r_{1y}
b_2z_2	$b_1b_2r_{21}$	b_2^2	b_2r_{2y}
z_y	b_1r_{y1}	b_2r_{y2}	1

Covariance $((x_1\ x_2)\ y) = b_1r_{y1} + b_2r_{y2}$

Variance $(x_1\ x_2) = b_1^2 + b_1b_2r_{21} + b_1b_2r_{21} + b_2^2$

Variance $(y) = 1$

$$b_1 = \frac{r_{1y} - r_{12} * r_{2y}}{1 - r^2_{12}} \qquad b_2 = \frac{r_{2y} - r_{12} * r_{1y}}{1 - r^2_{12}}$$

$$C_{12,y} = \frac{(r_{1y} - r_{12} * r_{2y}) * r_{1y} + (r_{2y} - r_{12} * r_{1y}) * r_{2y}}{1 - r^2_{12}}$$

$V_{12}=$

$$\frac{[(r_{1y} - r_{12} * r_{2y})^2 + 2 * (r_{1y} - r_{12} * r_{2y}) * (r_{2y} - r_{12} * r_{1y}) * r_{12} + (r_{2y} - r_{12} * r_{1y})^2}{(1 - r^2_{12})(1 - r^2_{12})}$$

expand and collect terms ==>

$$V_{12} = Cov = R^2_{12,y} = \frac{r^2_{1y} + r^2_{2y} - 2 * r_{1y} * r_{12} * r_{2y}}{1 - r^2_{12}}$$

if $r_{12} = 0$ then notice that $R^2_{12,y} = r_{1y}^2 + r_{2y}^2$

Unit Weights versus Multiple R

Multiple Correlation is optimally weighted composite:

$$r_{xy}=\frac{C_{xy}}{\sqrt{V_x V_y}}$$

But consider what happens if we equal (unit) weights rather than optimal weights

Standardized solution with unit weights

	z_1	z_2	z_y
z_1	1	r_{12}	r_{1y}
z_2	r_{21}	1	r_{2y}
z_y	r_{y1}	r_{y2}	1

Covariance $((x_1\ x_2)\ y) = r_{y1} + r_{y2}$

Variance $(x_1\ x_2) = 1 + r_{12} + r_{21} + 1$

Variance $(y) = 1$

$$R = \frac{r_{y1} + r_{y2}}{\sqrt{1 + r_{12} + r_{21} + 1}} \qquad = \frac{r_{y1} + r_{y2}}{\sqrt{2*(1 + r_{12})}}$$

Consider several examples:

rx1x2	rx1y	rx2y	beta 1	beta 2	R	R^2	Unit Wt	UW^2
0.0	0.5	0.5	0.50	0.50	0.71	0.50	0.71	0.50
0.3	0.5	0.5	0.38	0.38	0.62	0.38	0.62	0.38
0.5	0.5	0.5	0.33	0.33	0.58	0.33	0.58	0.33
0.7	0.5	0.5	0.29	0.29	0.54	0.29	0.54	0.29
0.3	0.5	0	0.55	-0.16	0.52	0.27	0.31	0.10
0.3	0.5	0.3	0.45	0.16	0.52	0.27	0.50	0.25

Partial Correlation

$$r_{xy}=\frac{C_{xy}}{\sqrt{V_x V_y}}$$

To find r_{xy} with w held constant (partial $r=r_{xy.w}$) or R_{xyw} (multiple R), we need to find the Covariance and Variances.

Conceptual solution:

find residual x after predicting from w (x.w)
find residual y after predicting from w (y.w)
correlate these residual scores.

Variance of residual = (Variance of original)*($1-r^2$)
Covariance of residuals =
original covariance - covariance with control

$z_{predicted} = r^*z_{predictor}$
$z_{residual} = z_{original} - r^*z_{predictor}$
$z_{x.w}=z_x-r_{xw}^*z_w$ $\qquad$ $z_{y.w}=z_y-r_{yw}^*z_w$

Covariance$(z_{x.w},z_{y.w})$=Cov(z_x,z_y) - $r_{xw}^*r_{yw}$ since
Covariance$(z_{x.w},z_{y.w})$=$\bullet(z_x-r_{xw}^*z_w)^*(z_y-r_{yw}^*z_w)/N$ =
$\bullet(z_x-r_{xw}^*z_w)^*(z_y-r_{yw}^*z_w)/N$ =
$\bullet(z_x^*z_y-r_{xw}^*z_w{}^*z_y - z_x{}^*r_{yw}^*z_w+r_{xw}^*z_w{}^*r_{yw}^*z_w)/N$ =
$\{\bullet(z_x{}^*z_y) - r_{xw}^*\bullet(z_w{}^*z_y) - r_{yw}^*\bullet z_x{}^*z_w+r_{xw}^*r_{yw}\bullet z_w{}^*z_w\}/N$
Cov(z_x,z_y) $-r_{xw}^*r_{wy}$ $- r_{yw}^*r_{xw}$ + $r_{xw}^*r_{yw}^*$Var z_w =
Cov(z_x,z_y) - $r_{xw}^*r_{yw}$
Variance residual = V_x * $(1-r_{xw}{}^2)$

$$\text{Partial } r_{xy.w} = \frac{\text{Cov}(z_x,z_y) - r_{xw}^*r_{yw}}{\sqrt{(1-r_{xw}{}^2)^*(1-r_{yw}{}^2)}}$$