Tinyurl.com/COMPEXC

(Half-)Detailed statistical companion for the 'Intro to stats in Excel' posted online

Content:

0. Preliminaries = "Chapter 0"

A. Tinyurl.com/INTRSTATS1 + Excel @ Tinyurl.com/101STATSEXCEL

1. RQ.1. Research Question 1: Are there more overweight males than females?

2. RQ.2. Research Question 2: Do these 10 cases change their BMI levels?

3. RQ.3. Research Question 1: Is the level of HgA1c predicted by BMI?

B. Tinyurl.com/INTRSTATS2 + Excel @ Tinyurl.com/101STATSEXCEL

4. RQ.4. Research Question : What is the effect of BMI on Systolic Blood Pressure (BP)?

5. RQ.4.a. Research Question : What is the effect of BMI on Systolic Blood Pressure (BP), net of (distinct from) that of HgA1c?

6. RQ.4.b. Research Question : Does the direct effect of BMI on Systolic Blood Pressure (BP), vary with the level of HgA1c?

7. RQ.4.c. Research Question : What is the total effect of BMI on Systolic Blood Pressure (BP), including the indirect effect through HgA1c?

8. RQ.4.d. Research Question : Does the indirect effect of BMI on Systolic Blood Pressure (BP), through HgA1c, vary with the very level of BMI?

9. RQ.5.: Is there a common cause (latent factor, e.g. "Bad Health") behind all 3: BMI, HgA1c, and Systolic BP?

10. RQ.6.: Are cases changing over 3 time points similarly? (Are those starting off higher changing slower?)

C. Spatial Tinyurl.com/INTRSTATS3 with Tinyurl.com/BLOGSTATS3 Excel @ Tinyurl.com/SPATIALSSM

11. RQ.7.: Are states with more residents in poverty expected to live shorter lives? By how much, comparing classic vs. spatial regression?

12. XXX12

~~~~Let's go:~~~~~~

0. Preliminaries = "Chapter 0"

* All these equations can be replicated by hand, or in some software, but they all come from existing textbooks, they are referenced.
* Nowadays however things are in flux: AI tools can be used to both figure out any question, any equation, any method: it just needs a bit of back and forth.
* There are three main statistical tests to be looked at; yes, many more exist, but we can run with these a long way: $t_{(df)}$ test (and z test, replacing it in many instances), $\chi^2_{(df)}$, and $F_{(df1, df2)}$
* Much confusion when learning statistical tricks comes from fluid notation and labeling, which loses meaning along the way; here 2 solutions are offered:
i. Preserving meaning and scaling, like variance is a 'squared kilograms' (for Y = weight in kilograms, e.g.) so we label it σ_y^2 vs. standard deviation is a 'kilograms' quantity therefore labeled σ_y
ii. Replacing the Greek with plain English like σ_y^2 = var(y) and σ_y = sd(y), and σ_{xy}^2 = cov(x,y) and therefore $\sigma_{xy} = \overset{notation}{\sim} \rho$ = cov(x,y)
* 'Doing the math' can be pursued in several ways, all arriving at the same results:
(1). By algebraic means, writing equations, and multiplying on both sides, adding up, etc.;
~~~~

(2). By means of vectors and their older relatives matrices: this simplifies things, but require a bit of intro; **e.g.**
Written as vector, variance (of DEVIATIONS) $\sigma_x^2 = \Sigma_i x_i \cdot x_i = \mathbf{x}' \cdot \mathbf{x} = (x_1, \ldots, x_i, \ldots, x_N) \cdot [x1 \ \ldots \ xi \ \ldots \ xN]$
(3). By an extension of algebra, which uses a few emergent 'covariance algebra' rules;
(4). Using a graphical method of variables as vertices/dots and relations as arrows – the 'tracing rule' (we will rely a bunch on this);
(5). Possibly also by an extension of the above, into geometry (fun, but takes a longer 'warm up'.
(6). Practically software algorithms utilize several minimization or maximization algorithms to arrive at a 'best value' - these algorithms often work where no 'closed form algebraic formula' can be really worked out.
X. [+ maybe more than this]

1. RQ.1. Research Question 1: Are there more overweight males than females?

Model depiction: Female/Male -> Overweight/Not

1. The $\chi^2_{(df)}$ formula is implemented simply in Excel through 'CHITEST(C32:D33,I32:J33)', as an example, where C32:D33 and I32:J33 are 2 identical 'blocks': actual_range and expected_range; it simply implements the is: $\chi^2 = \sum(O_i - E_i)^2/E_i$, where O_i = observed value (actual value) and E_i = expected value, and hence it has the same units as the original values (kilograms, e.g.), because it's a kilograms^2/kilograms quantity
2. The larger the value, the more departure of observed values from expected. How big the departure needs to be to conclude 'they cannot be the same values (as 2 groups, 2 sets! not individual values) is what the p values helps one decide, e.g. looking online WWW1.
3. The Excel setup makes 'visible' what degrees of freedom are! For an existing setup, meanin the row totals and column totals are known (=fixed), once we set 1 number inside the table, all the other 3 (in a 2x2 table) are determined! Hence df = 1 !!! Truly only 1 way in which the data - as 4 aggregates in a 2x2 table – has only 1 freedom-dimension, or ability to truly 'vary' freely: the others are not free anymore after that!
4.
GO BACK UP

2. RQ.2. Research Question 2: Do these 10 cases change their BMI levels?

Model depiction: Female/Male □ Overweight/Not

See 4. RQ.4.

GO BACK UP

3. RQ.3. Research Question 1: Is the level of HgA1c predicted by BMI?

Model depiction: Overweight/Not.W1 □ Overweight/Not.W2

See 4. RQ.4.

GO BACK UP

4. RQ.4. Research Question : What is the effect of BMI on Systolic Blood Pressure (BP)?

Model depiction: BMI.cont □ SysBP.cont

1. Classic algebra Greene_2002_Econometric Analysis_BOOK

Now divide both the numerator and denominator in the expression for b_3 by $\Sigma_i t_i^2 \Sigma_i g_i^2$. By manipulating it a bit and using the definition of the sample correlation between G and T, $r_{gt}^2 = (\Sigma_i g_i t_i)^2/(\Sigma_i g_i^2 \Sigma_i t_i^2)$, and defining b_{yt} and b_{tg} likewise, we obtain

$$b_{yg\cdot t} = \frac{b_{yg}}{1 - r_{gt}^2} - \frac{b_{yt} b_{tg}}{1 - r_{gt}^2} = 0.653723. \qquad \textbf{(3-10)}$$

(The notation "$b_{yg\cdot t}$" used on the left-hand side is interpreted to mean the slope in the regression of y on g "in the presence of t.") The slope in the **multiple regression** differs from that in the simple regression by including a correction that accounts for the influence of the additional variable t on both Y and G. For a striking example of this

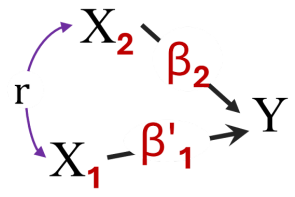
b.

It is thus seen that any partial correlation coefficient may be expressed in terms of the central moments λ_{ik}, or the total correlation coefficients ϱ_{ik} of the variables concerned. Thus we obtain, e. g., in the case $n = 3$

(23.4.3) $$\varrho_{12\cdot 3} = \frac{\varrho_{12} - \varrho_{13}\varrho_{23}}{\sqrt{(1 - \varrho_{13}^2)(1 - \varrho_{23}^2)}}.$$

Harald Cramer 1946
Mathematical Methods of Statistics p. 306

2. Tracing rule – notation change:
$b_{yx1\cdot x2} = (b_{yx1} - b_{yx2} \cdot b_{x1x1}) / (1 - r^2_{x1x2})$
$r_1 = \text{corr}(X_1,Y)$ $r_2 = \text{corr}(X_2,Y)$ $r = \text{corr}(X_1,X_2)$
Absent X_2, $\beta_1 = r_1$
With X_2,
$r_1 = \beta_1 + r \cdot \beta_2$,
$r_2 = \beta_2 + r \cdot \beta_1$, hence $\beta_2 = r_2 - r \cdot \beta_1$
so $\beta_1 = r_1 - r \cdot \beta_2 = r_1 - r \cdot r_2 - r^2 \cdot \beta_1$
$\beta_1 \cdot (r^2 - 1) = = r_1 - r * r_2$ so:
$\beta_1 = (r_1 - r * r_2) / (1 - r^2)$
By symmetry then
$\beta_2 = (r_2 - r * r_1) / (1 - r^2)$
[$b_{yx1\cdot x2} = (b_{yx1} - b_{yx2} \cdot b_{x1x1}) / (1 - r^2_{x1x2})$!]



GO BACK UP

5. RQ.4.a. Research Question : What is the effect of BMI on Systolic Blood Pressure (BP), net of (distinct from) that of HgA1c?

Model depiction: BMI.cont & HgA1c □ SysBP.cont

1. Classic algebra

2. Tracing rule

GO BACK UP

6. RQ.4.b. Research Question : Does the direct effect of BMI on Systolic Blood Pressure (BP), vary with the level of HgA1c?

Model depiction: BMI & HgA1c & BMI *HgA1c □ SysBP

1. Classic algebra

2. Tracing rule

GO BACK UP

7. RQ.4.c. Research Question : What is the total effect of BMI on Systolic Blood Pressure (BP), including the indirect effect through HgA1c?

Model depiction: BMI □ HgA1c □ SysBP + BMI □ SysBP

1. Classic algebra

2. Tracing rule

GO BACK UP

8. RQ.4.d. Research Question : Does the indirect effect of BMI on Systolic Blood Pressure (BP), through HgA1c, vary with the very level of BMI?

Model depiction: BMI □ HgA1c □ SysBP + BMI & BMI. *HgA1c □ SysBP

1. Classic algebra

2. Tracing rule

GO BACK UP

9. RQ.5.: Is there a common cause (latent factor, e.g. "Bad Health") behind all 3: BMI, HgA1c, and Systolic BP?

Model depiction: LATENT*BadHealth*.cont □ [BMI & HgA1c & SysBP]

Exhibit A3.2 Covariance algebra rules

Defining c as a constant and X_1, X_2 and X_3 as random variables (see *Note*),

1 $\text{cov}(c, X_1) = 0$. (The covariance between a constant and a random variable is 0.)

2 $\text{cov}(cX_1, X_2) = c\text{cov}(X_1, X_2)$.

3 $\text{cov}(X_1 + X_2, X_3) = \text{cov}(X_1, X_3) + \text{cov}(X_2, X_3)$.

4 $\text{cov}(X_1, X_1) = \text{var}(X_1)$. (The covariance of a variable with itself is its variance.)

McCoach, D. B., & Cintron, D. (2022). Introduction to modern modelling methods [app. Tinyurl.com/CINTRONAPP]

These rules are derivable by writing case level equations and multiplying on both sides, summing up, and dividing by M

!!! Note that $\text{cov}(y_1 \cdot y_2, x)$ is not easy to carve out!

1. Covariance algebra – classic algebra is parallel proof, just multiply each y, e.g. y_{1i} by another y, e.g. y_{2i}, sum up across I, and divide by N

$y_{1i} = \alpha_1 + \lambda_1 \cdot L + \varepsilon_{1i}$ L is Latent variable – not in the data! Like ε's!

$y_{2i} = \alpha_2 + \lambda_2 \cdot L + \varepsilon_{2i}$

$y_{3i} = \alpha_3 + \lambda_3 \cdot L + \varepsilon_{3i}$

Show for y_1 and y_2 :

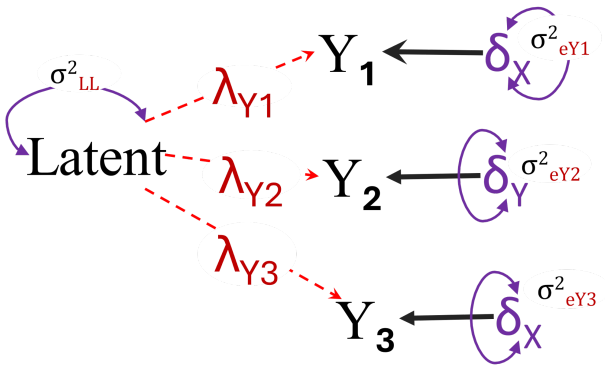
$\text{cov}(y_1, y_2) = \text{cov}(\alpha_1 + \lambda_1 \cdot L + \varepsilon_{1i}, \alpha_2 + \lambda_2 \cdot L + \varepsilon_{2i}) =$

$+ \textit{cov}(\alpha_1, \alpha_2) + \textit{cov}(\alpha_1, \lambda_2 \cdot L) + \textit{cov}(\alpha_1, \varepsilon_{2i}) +$

$+ \textit{cov}(\lambda_1 \cdot L, \alpha_2) + \text{cov}(\lambda_1 \cdot L, \lambda_2 \cdot L) + \underline{\text{cov}(\lambda_1 \cdot L, \varepsilon_{2i})} +$

$\textit{cov}(\varepsilon_{1i}, \alpha_2) + \text{cov}(\varepsilon_{1i}, \lambda_2 \cdot L) + \underline{\text{cov}(\varepsilon_{1i}, \varepsilon_{2i})}$

And all italics are 0 because one term is a constant; $\text{cov}(\varepsilon_{1i}, \varepsilon_{2i})$ @=0 assumed, can be relaxed; $\text{cov}(\lambda_1 \cdot L, \varepsilon_{2i})$ & $\text{cov}(\varepsilon_{1i}, \lambda_2 \cdot L)$ @=0 assumed, and $\text{cov}(\lambda_1 \cdot L, \lambda_2 \cdot L) = \lambda_1 \cdot \lambda_2 \cdot \text{cov}(L, L) = \lambda_1 \cdot \lambda_2 \cdot \text{var}(L) = \lambda_1 \cdot \lambda_2$ because is var(L) is ASSUMED =1: this identifying condition is less common, normally folks know about one loading being set @1, like , but this is problematic, changing it changes the estimates, among others. The math works anyways. So:



$\text{cov}(y_1, y_2) = \lambda_1 \cdot \lambda_2$, similarly by symmetry

$\text{cov}(y_1, y_3) = \lambda_1 \cdot \lambda_3$,

$\text{cov}(y_2, y_3) = \lambda_2 \cdot \lambda_3$, identical solution to below; 3 equations with 3 unknowns, hence λ's can be calculated.

$\lambda_{Y1} = \text{SQRT}(\text{cov}(y_1, y_2) \cdot \text{cov}(y_1, y_3) / \text{cov}(y_2, y_3))$

$\lambda_{Y2} = \text{SQRT}(\text{cov}(y_1, y_2) \cdot \text{cov}(y_2, y_3) / \text{cov}(y_1, y_3))$

$\lambda_{Y3} = \text{SQRT}(\text{cov}(y_2, y_3) \cdot \text{cov}(y_1, y_3) / \text{cov}(y_1, y_2))$

2. Tracing rule

$\text{cov}(Y_1, Y_2) = \sigma^2_{Y1Y2} = \lambda_{Y1} \cdot \lambda_{Y2}$ *[trek Y_1 -> L-> Y_2]*

$\text{cov}(Y_1, Y_3) = \sigma^2_{Y1Y3} = \lambda_{Y1} \cdot \lambda_{Y3}$ *[trek Y_1 -> L-> Y_3]*

$\text{cov}(Y_2, Y_3) = \sigma^2_{Y2Y3} = \lambda_{Y2} \cdot \lambda_{Y3}$ *[trek Y_2 -> L-> Y_3]*

Solving for λ's we get

For the % unreliability (measurement error):

$\text{cov}(Y_1, Y_1) = \text{var}(Y_1) = \sigma^2_{Y1Y1} = = \lambda_{Y1} \cdot \lambda_{Y1} + \sigma^2_{eY1}$

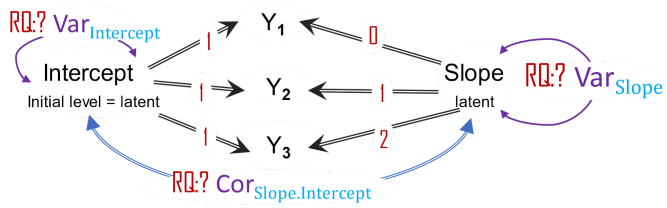
[trek Y_1 -> L-> Y_1 , & Y_1 -> e_{Y1} -> Y_1]
Variance = Explained + Unexplained
= Due to Latent(+RELIABLE) + Unreliability

$\sigma^2_{eY1} = \sigma^2_{Y1} - \sigma^2_{Y1Y2} \cdot \sigma^2_{Y1Y3} / \sigma^2_{Y2Y3}$
$\sigma^2_{eY2} = \sigma^2_{Y2} - \sigma^2_{Y1Y2} \cdot \sigma^2_{Y2Y3} / \sigma^2_{Y1Y3}$
$\sigma^2_{eY3} = \sigma^2_{Y3} - \sigma^2_{Y2Y3} \cdot \sigma^2_{X1Y3} / \sigma^2_{Y1Y2}$

GO BACK UP

10. RQ.6.: Are cases changing over 3 time points similarly? (Are those starting off higher changing slower?)

This is a PRETEND setup; new data can be easily generated BUT it's used this way to show the IDENTITY between the data in the 3 variable (factor analysis e/.g.) setup and the 1 variable 3 time points (changes) designs.



- Li is latent intercept, and Ls latent slope: NOT in the data!

Model depiction: Li & 0*Ls □ Y1 + Li & 1*Ls □ Y2 + Li & 2*Ls □ Y1

1. Covariance algebra – classic algebra is parallel proof, just multiply each y, e.g. y_{1i} by another y, e.g. y_{2i}, sum up across I, and divide by N

$y_{1i} = \alpha_1 + 1 \cdot L_I + 0 \cdot L_S + \varepsilon_{1i}$ L is Latent variable – not in the data! Like ε's!

$y_{2i} = \alpha_2 + 1 \cdot L_I + 1 \cdot L_S + \varepsilon_{2i}$

$y_{3i} = \alpha_3 + 1 \cdot L_I + 2 \cdot L_S + \varepsilon_{3i}$

Show for y_1 and y_2 :

$\text{cov}(y_1 , y_2) = \text{cov}(\alpha_1 + 1 \cdot L_I + 0 \cdot L_S + \varepsilon_{1i} , \alpha_2 + 1 \cdot L_I + 1 \cdot L_S + \varepsilon_{2i}) =$ [all terms with α's and ε's are 0] =

$= cov(L_I, L_I) + cov(L_I, 1 \cdot L_S) =$

$= var(L_I) + cov(L_I, L_S)$

$\text{cov}(y_1 , y_3) = \text{cov}(\alpha_1 + 1 \cdot L_I + 0 \cdot L_S + \varepsilon_{1i} , \alpha_3 + 1 \cdot L_I + 2 \cdot L_S + \varepsilon_{3i}) =$ [all terms with α's and ε's are 0] =

$= cov(L_I, L_I) + cov(L_I, 2 \cdot L_S) =$

$= var(L_I) + 2 \cdot cov(L_I, L_S)$

$\text{cov}(y_2 , y_3) = \text{cov}(\alpha_2 + 1 \cdot L_I + 1 \cdot L_S + \varepsilon_{2i} , \alpha_3 + 1 \cdot L_I + 2 \cdot L_S + \varepsilon_{3i}) =$ [all terms with α's and ε's are 0] =

$= cov(L_I, L_I) + cov(L_I, 2 \cdot L_S) + cov(1 \cdot L_S, L_I) + cov(1 \cdot L_S, 2 \cdot L_S) =$

$= var(L_I) + \mathbf{3} \cdot cov(L_I, L_S) + 2 \cdot var(L_S)$

And then we can solve for the 3: $var(L_S)$ & $var(L_I)$ & $cov(L_I, L_S)$

2. Tracing rule – the **2** does NOT match !

$\text{cov}(Y_1,Y_2) = 1 \cdot \text{var}(L_I) \cdot 1 + 1 \cdot \text{cov}(L_I,L_S) \cdot 1 + 0 \text{ var}(L_S) \cdot 1 = \text{var}(L_I) + \text{cov}(L_I,L_S)$

$\text{cov}(Y_1,Y_3) = 1 \cdot \text{var}(L_I) \cdot 1 + 1 \cdot \text{cov}(L_I,L_S) \cdot 2 + 0 \cdot \text{var}(L_S) \cdot 2 = \text{var}(L_I) + 2 \text{ cov}(L_I,L_S)$

$\text{cov}(Y_2,Y_3) = 1 \cdot \text{var}(L_I) \cdot 1 + 1 \cdot \text{cov}(L_I,L_S) \cdot 2 + 1 \cdot \text{var}(L_S) \cdot 2 = \text{var}(L_I) + \mathbf{2} \text{ cov}(L_I,L_S) + 2\text{var}(L_S)$

Therefore one can get the 3: $\text{var}(L_I)$ & $\text{var}(L_S)$ & $\text{cov}(L_I,L_S)$

1 and 2 without $\text{var}(L_I)$: $\text{cov}(Y_1,Y_2) - \text{cov}(L_I,L_S) = \text{cov}(Y_1,Y_3) - 2 \cdot \text{cov}(L_I,L_S)$

so $\text{cov}(L_I,L_S) = \text{cov}(Y1,Y3) - \text{cov}(Y1,Y2)$ [does match LGM Stata estimate!]

Back into 1: $\text{var}(L_I) = 2 \cdot \text{cov}(Y_1,Y_2) - \text{cov}(Y_1,Y_3)$

[does match LGM Stata estimate!]

Now in 3: $\text{var}(L_S) = \text{cov}(Y_2Y_3) - \text{var}(L_I) - 2 \cdot \text{cov}(L_I,L_S) =$

$(\text{cov}(Y_2,Y_3) - 2 \cdot \text{cov}(Y_1,Y_2) + \text{cov}(Y_1,Y_3) - 2 \cdot \text{cov}(Y_1,Y_3 + 2 \cdot \text{cov}(Y_1,Y_2) / 2$

$= (\text{cov}(Y_2,Y_3 - \text{cov}(Y_1,Y_3) / 2$

GO BACK UP

11. RQ.7.: Are states with more residents in poverty expected to live shorter lives? By how much, comparing classic vs. spatial regression?

Model depiction: $\%\text{Poverty}_{\text{USStates}}$ □ $\text{LifeExpectancy}_{\text{USStates}}$

1. The

GO BACK UP

12. xxxxxxxxxxxxxxxx

Model depiction: XXX □ YYY

GO BACK UP