Statistics 2

Multiple regression: Partial Correlation. Standardized Regression

Casper Albers & Jorge Tendeiro

Lecture 6, 2019 – 2020

university of groningen

Overview

Partial correlation
Semi-partial correlation
Ballantine Venn diagrams
Squared semipartial correlation
Squared partial correlation
Standardized regression coefficients

Literature for this lecture

Read:
Agresti, Section 11.6 - 11.7

Note: The material on Ballantine Venn diagrams and the semi-partial correlation on the slides is not part of the book but it is part of the exam material!

Example – Predicting academic performance

Niessen et al. (2016) studied various predictors of academic performance[1].

- Dependent variable
  - y: FYGPA, grade point average of courses in the first year.
- Independent variables ($p = 2$)
  - x_1: CST, curriculum-sampling test; a test representative for the psychology curriculum.
  - x_2: HSGPA, the grade point average obtained in high school.

Correlations ($n = 201$):

	y	x_1	x_2
y			
x_1	.451		
x_2	.408	.523	

How much is the correlation between y and x_i when controlling for x_j?

[1] Niessen, A.S.M., Meijer, R.R., Tendeiro, J.N. (2016). Predicting Performance in Higher Education Using Proximal Predictors. *PLoS ONE*, doi:10.1371/journal.pone.0153663

Partial correlation

In case of $k = 2$ predictors:

The partial correlation between y and x_1 is the correlation of 'y without x_2' with 'x_1 without x_2'.

$$pr_1 = r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1 - r^2_{yx_2})(1 - r^2_{x_1x_2})}}.$$

$$pr_1 = r_{yx_1 \cdot x_2} = \frac{.451 - .408 \times .523}{\sqrt{(1 - .408^2) \times (1 - .523^2)}} = \frac{.238}{.778} = .305$$

The correlation between average first year grade (y) and curriculum sampling test (x_1), when partialing out the effect of the average high school grade, is .305.

Similarly,

$$pr_2 = r_{yx_2 \cdot x_1} = \frac{.408 - .451 \times .523}{\sqrt{(1 - .451^2) \times (1 - .523^2)}} = \frac{.172}{.579} = .297.$$

Alternative approach ($k = 2$)

1. Partial out x_2 from y:

$$y_i = \alpha_0 + \alpha_1 x_{2,i} + e_{y,i}$$

   Here, e_y are what is left over from y once x_2 is taken out;

2. Partial out x_2 from x_1:

$$x_{i,1} = \gamma_0 + \gamma_1 x_{2,i} + e_{x_1,i}$$

   Here, e_{x1} is what is left over from x_1 once x_2 is taken out;

3. Compute the regular (zero-order) correlation between e_y and e_{x_1}:

$$pr_1 = r_{yx_1 \cdot x_2} = \text{cor}(e_y, e_{x_1}).$$

Alternative approach ($k = 2$) – Example

1. Partial out x_2 from y:

$$y = \alpha_0 + \alpha_1 x_2 + e_y$$

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	3.549	.472		7.515	.000
	CST	.099	.016	.408	6.311	.000

2. Partial out x_2 from x_1:

$$x_{i,1} = \gamma_0 + \gamma_1 x_{2,i} + e_{x_1,i}$$

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	−2.134	.999		−2.136	.034
	HSGPA	1.295	.150	.523	8.655	.000

3. For both analyses, save the residuals and compute the correlation:

$$pr_1 = .305$$

Partial correlation when $k > 2$

In case of $k > 2$ predictors:

The partial correlation between y and x_1 is the correlation of 'y without $x_2, x_3, \ldots$' with 'x_1 without $x_2, x_3, \ldots$'.

Computation:

Can be done with direct formula (see p. 346, not exam material), or indirect approach.

Alternative approach ($k > 2$)

Partial correlations can be computed for every pair of variables. For instance $\{x_1, y\}$:

1. Partial out $x_2, \dots, x_p$ from y:

$$y_i = \alpha_0 + \alpha_1 x_{2,i} + \dots + \alpha_{p-1} x_{p,i} + e_{y,i}$$

   Here, e_y is what is left over from y once all x but x_1 are taken out;

2. Partial out $x_2, \dots, x_p$ from x_1:

$$x_{i,1} = \gamma_0 + \gamma_1 x_{2,i} + \dots + \gamma_{p-1} x_{p,i} + e_{x_1,i}$$

   Here, e_{x1} is what is left over from x_1 once all x but x_1 are taken out;

3. Compute the regular (zero-order) correlation between e_y and e_{x_1}:

$$r_{yx_1 \cdot x_2, \dots, x_p} = \mathrm{cor}(e_y, e_{x_1}).$$

Semi-partial correlation

- The partial correlation removes $x_2, x_3, \ldots$ from y and x_1 and then computes the correlation.
- The semi-partial correlation removes $x_2, x_3, \ldots$ from x_1 only and then computes the correlation.

Computation:

1. Do nothing with y
2. Partial out $x_2, \ldots, x_p$ from x_1:

$$x_{i,1} = \gamma_0 + \gamma_1 x_{2,i} + \ldots + \gamma_{p-1} x_{p,i} + e_{x_1,i}$$

   Here, e_{x1} is what is left over from x_1 once all x but x_1 are taken out;
3. Compute the regular (zero-order) correlation between y and e_{x_1}:

$$sr_1 = \mathrm{cor}(y, e_{x_1}).$$

Semi-partial correlation – direct computation

The following formula gives the same result for $k = 2$ predictors:

$$sr_1 = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{1 - r^2_{x_1x_2}}}$$

Example:

$$sr_1 = \frac{.451 - .408 \times .523}{\sqrt{1 - .523^2}} = \frac{.238}{.852} = .279,$$
$$sr_2 = \frac{.408 - .451 \times .523}{\sqrt{1 - .523^2}} = \frac{.172}{.852} = .202.$$

For $k > 2$ direct computation formulas do exist, but are not part of the course material.

Interpretation squared correlation

- Pearson correlation, r, measures strength (and direction) of relation between x and y.
- r^2: percentage of variance in common by x and y.
- partial correlation, pr, and semi-partial correlation, sr, have interpretation as a special type of correlation.
- Do their squared values, pr^2 and sr^2 have a 'percentage explained variance'-interpretation?

Answer: Yes.

To see this, we use the Ballantine Venn Diagram.

Example

- Dependent variable: y = FYGPA
- Independent variables: x_1: CST and x_2: HSGPA

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	-2.085	.976		-2.137	.034
	CST	.052	.016	.217	3.277	.001
	HSGPA	1.053	.164	.425	6.434	.000

Coefficients

Model	R	R Square	Adjusted R Square	Std. Error of the estimate
1	.558	.311	.304	.913

Important: The univariate R^2's don't add up to the multiple R^2:

$$R^2 \neq r_{y1}^2 + r_{y2}^2.$$

Example

Simple Linear Regression

Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.408[a]	.167	.163	1.001

[a] Predictors: (Constant), CST

Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.523[a]	.273	.270	.935

[a] Predictors: (Constant), HSGPA

Multiple Regression

Model summary

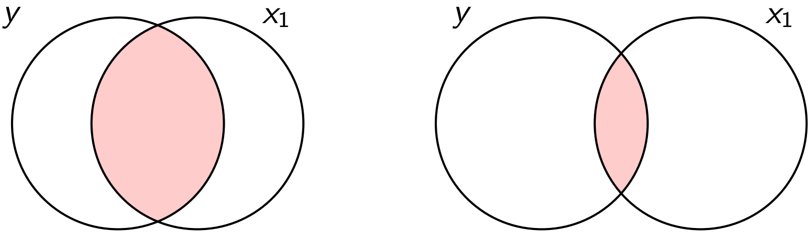
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.558[a]	.311	.304	.913

[a] Predictors: (Constant), CST, HSGPA

$$R^2 \neq r_{y1}^2 + r_{y2}^2.$$

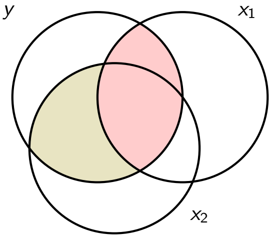
$$.311 \neq .167 + .273 = .440.$$

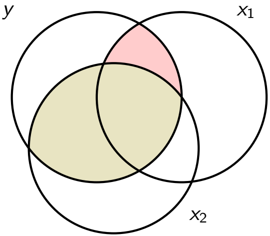
Ballentine Venn Diagram



- Conceptual method of visualizing strength of relations between variables.
- The larger the overlap between the circles for y and x_1, the larger the explained variance ($r^2_{yx_1}$).
- Left: Overlap, r^2, 53%. Right: Approximately 14%.

Ballentine Venn Diagram





$r^2_{yx_1} = 0.167$

$R^2 - r^2_{yx_1} = 0.144$

$r^2_{yx_2} = 0.273$

$R^2 - r^2_{yx_2} = 0.038$

Adding $x_1 = \text{CST}$ when $x_2 = \text{HSGPA}$ is already in the model:
R^2 barely increases.

Explanation

The univariate R^2's (usually) don't add up to the multiple R^2.

Explanation:

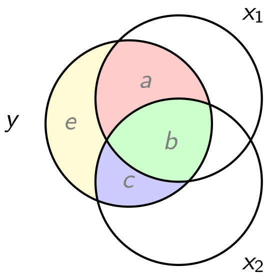
- Multicollinearity:
  Correlation between predictor variables.
- CST explains 17% of variance in FYGPA and HSGPA explains 27%.
- Together not 44% because of the large overlap between CST and HSGPA: $r_{x_1 x_2} = .523$.
- Only if $r_{x_1 x_2} = 0$ (completely uncorrelated predictor variables) then $R^2 = r^2_{yx_1} + r^2_{yx_2}$.

Multicollinearity

Multicollinearity: Correlation between independent variables.

- Consequences of large correlation:
  - Difficult to decide which variables are 'important'.
  - Large standard errors.
- How to check for multicollinearity:
  - Inspect the bivariate correlations between the independent variables.
  - The so-called Variance Inflation Factor measures the amount of multicollinearity.
  - Rule of thumb: $\text{VIF}_j < 4$ is okay. (Computation of VIF: Statistics III)
- Solution if multicollinearity is detected:
  - Using other variables, combining variables, not using (some) variables.

Ballentine Venn Diagram



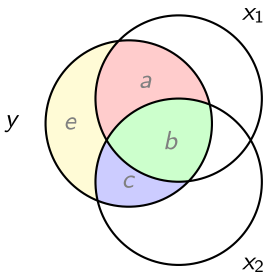
What do all the parts of the Ballantine mean?

- a: the *unique* contribution of x_1 to y.
- c: the *unique* contribution of x_2 to y.
- b: the *common* contribution of x_1 and x_2 to y.
- e: 'error': unexplained part of y.

$$r^2_{yx_1} = a + b \quad \text{and} \quad r^2_{yx_2} = c + b$$

$$R^2 = a + b + c$$

Squared semipartial correlation

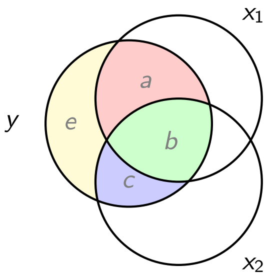


Semipartial correlation^2 of x_i: *How much of the total variance of y is uniquely explained by this IV?*

For x_1:

- $sr_1^2 = a$
- $sr_1^2 = R^2 - r_{yx_2}^2 = (a + b + c) - (b + c)$

Squared partial correlations



Partial correlation^2 of x_i: *What proportion of the variance of y not explained by the other IVs, is uniquely explained by this IV?*

For x_1:

- $pr_1^2 = \frac{a}{a+e} = \frac{a}{1-b-c}$
- $pr_1^2 = \frac{R^2 - r_{yx_2}^2}{1 - r_{yx_2}^2}$

Standardized regression coefficients

Up to now: Correlations.

Now let's focus on the regression coefficients.

Coefficients

Model		Unstandardized	Standard Error	Standardized	*t*	*p*
1	(Intercept)	-2.085	.976		-2.137	.034
	CST	.052	.016	.217	3.277	.001
	HSGPA	1.053	.164	.425	6.434	.000

How to interpret, e.g., 0.052?

- If HSGPA is kept fixed, then each increase by 1 unit in CST represents a .052 increase in the mean of FYGPA.
- How to understand '1 unit increase in CST'? Not possible with background knowledge of study.
- To overcome this: Standardize all variables and re-run the regression.

Standardized regression coefficients

Original regression model:

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + e_i.$$

Regression model based on standardized values:

$$y_i^* = b_0^* + b_1^* x_{1,i}^* + b_2^* x_{2,i}^* + e_i^*.$$

here, y^* and x_i^* are standardized (mean 0, sd 1) and

$$b_1^* = b_1 \frac{s_{x_1}}{s_y}, \quad b_1^* = b_2 \frac{s_{x_2}}{s_y}.$$

Standardized regression coefficients

Coefficients

Model		Unstandardized	Standard Error	Standardized	*t*	*p*
1	(Intercept)	-2.085	.976		-2.137	.034
	CST	.052	.016	.217	3.277	.001
	HSGPA	1.053	.164	.425	6.434	.000

$$y_i^* = b_0^* + .217x_{1,i}^* + .425x_{2,i}^* + e_i^*.$$

Easier to interpret:

1 SD increase in CST, when holding HSGPA constant, leads to .217 SD increase in mean FYGPA.

Next week: Assumptions

- Assumptions

Read: Only the lecture slides