Formulas

Stephanie Ranft S2459825

November 4, 2019

Statistics 2
PSBE2-07

Exercises

Semi-partial and partial correlation

Description:

This fictional data set, "Exam Anxiety", provides questionnaire scores by students prior to an exam (the variables are anxiety, preparedness, and grade).

`Revise` Time spent studying for the exam (in hours).

`Exam` Performance in the exam (percentages).

`Anxiety` Anxiety prior to the exam as measured by the Exam Anxiety Questionnaire.

(a) Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.457	0.209	0.193	23.306

(b) ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	14321.514	2	7160.757	13.184	< .001
	Residual	54315.690	100	543.157		
	Total	68637.204	102			

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	87.833	17.047		5.152	< .001
	Revise	0.241	0.180	0.169	1.339	0.184
	Anxiety	-0.485	0.191	-0.321	-2.545	0.012

(c) Coefficients

Table 1

We wish to know which independent variable (hours of revision or anxiety score prior to the exam) best describes the dependent variable (exam performance).

For convenience, Y = `exam`, X_1 = `revise`, and X_2 = `anxiety`, then $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ is the population regression equation. In order to compute the partial and semi-partial correlations, we need *other* regression equations. First, we regress X_1 and X_2 (separately) on Y, and also regress X_1 and X_2 on each other:

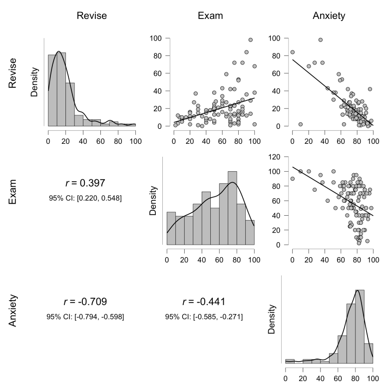

Figure 1

$$Y = \alpha_0 + \alpha_1 X_1 + e_{Y.X_1}; \qquad Y = \gamma_0 + \gamma_1 X_2 + e_{Y.X_2};$$
$$X_1 = \delta_0 + \delta_1 X_2 + e_{X_1.X_2}; \qquad X_2 = \kappa_0 + \kappa_1 X_1 + e_{X_2.X_1}.$$

Then we can calculate the partial correlation coefficients as

$$pr_1 = \mathrm{Cor}\left(e_{Y.X_2}, e_{X_1.X_2}\right) = \frac{r_{Y,X_1} - r_{Y,X_2} \cdot r_{X_1,X_2}}{\sqrt{\left(1 - r^2_{Y,X_2}\right) \cdot \left(1 - r^2_{X_1,X_2}\right)}}$$

$$pr_2 = \mathrm{Cor}\left(e_{Y.X_1}, e_{X_2.X_1}\right) = \frac{r_{Y,X_2} - r_{Y,X_1} \cdot r_{X_1,X_2}}{\sqrt{\left(1 - r^2_{Y,X_1}\right) \cdot \left(1 - r^2_{X_1,X_2}\right)}}$$

Using fig. 1, calculate the partial correlation coefficients for `revise` and `anxiety`. Describe what these values represent.

The semi-partial correlation coefficients are calculated as

$$sr_1 = pr_1 \cdot \sqrt{1 - r^2_{Y,X_2}} = \frac{r_{Y,X_1} - r_{Y,X_2} \cdot r_{X_1,X_2}}{\sqrt{1 - r^2_{X_1,X_2}}}; \qquad sr_2 = pr_2 \cdot \sqrt{1 - r^2_{Y,X_1}} = \frac{r_{Y,X_2} - r_{Y,X_1} \cdot r_{X_1,X_2}}{\sqrt{1 - r^2_{X_1,X_2}}}$$

Using fig. 1, calculate the semi-partial correlation coefficients for `revise` and `anxiety`. Describe what these values represent.

If we square the partial and semi-partial correlation coefficients, we can use the Ballentine Venn Diagram to compute their values.

$$pr_1^2 = \frac{R^2 - r^2_{Y,X_2}}{1 - r^2_{Y,X_2}}; \qquad pr_2^2 = \frac{R^2 - r^2_{Y,X_1}}{1 - r^2_{Y,X_1}};$$

$$sr_1^2 = pr_1^2 \cdot \left(1 - r^2_{Y,X_2}\right) = R^2 - r^2_{Y,X_2}; \qquad sr_2^2 = pr_2^2 \cdot \left(1 - r^2_{Y,X_1}\right) = R^2 - r^2_{Y,X_1}.$$

The circles represent the variances of each variable, and how they overlap. As none of our variables are uncorrelated, all circles overlap.

In words, what do the letters *a, b, c, e* represent?
Calculate the values of the squared partial and semi-partial correlations, and explain their meaning.
Use the values *a, b, c, e* to represent the squared partial and semi-partial correlations.
Which is a better predictor of the variable `exam`, `revision` or `anxiety`? Summarise your findings in a one or two sentences.

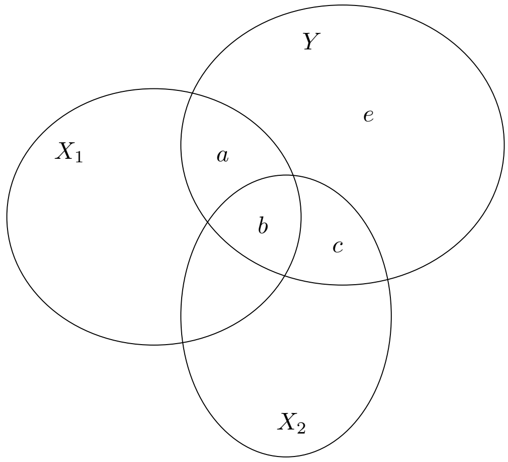

Figure 2

Solution.

$$pr_1 = \frac{0.397 - (-0.441) \cdot (-0.709)}{\sqrt{\left(1 - (-0.441)^2\right) \cdot \left(1 - (-0.709)^2\right)}} = 0.133$$

$$pr_2 = \frac{-0.441 - (0.397) \cdot (-0.709)}{\sqrt{\left(1 - (0.397)^2\right) \cdot \left(1 - (-0.709)^2\right)}} = -0.247$$

When partialing out the effect of anxiety on the other two variables, the correlation between exam results and the time spent revising is 0.133. Similarly, the correlation between exam results and pre-exam anxiety scores is -0.247 after partialing out the effect of revising on the other two variables.

$$sr_1 = 0.133 \cdot \sqrt{1 - (-0.441)^2} = \frac{-0.441 - (0.397) \cdot (-0.709)}{\sqrt{1 - (-0.709)^2}} = 0.199$$

$$sr_2 = -0.247 \cdot \sqrt{1 - (0.397)^2} = \frac{-0.441 - (0.397) \cdot (-0.709)}{\sqrt{1 - (-0.709)^2}} = -0.226$$

After partialling out the effect of anxiety on the time spent revising, the correlation of revision with exam results is 0.199. Similarly, the correlation between exam results and the part of anxiety unexplained by revision is -0.226.

a and b are the portions of the variance of Y which are solely explained by X_1 and X_2, respectively. b is the portion of the variance of Y explained dually by X_1 and X_2. e is the portion of the variance of Y which is not explained by X_1 nor X_2. Therefore, $R^2 = a + b + c$ is the variance of Y accounted for by the regression (VAF) and $e = 1 - R^2$ is the fraction of variance unaccounted for by the regression (FVU).

$$pr_1^2 = \frac{0.209 - (-0.441)^2}{1 - (-0.441)^2} = 0.0180 = \frac{a}{a+e}; \qquad pr_2^2 = \frac{0.209 - (0.397)^2}{1 - (0.397)^2} = 0.0610 = \frac{c}{c+e};$$

$$sr_1^2 = 0.0177 \cdot \left(1 - (-0.441)^2\right) = 0.0145 = a; \qquad sr_2^2 = 0.0610 \cdot \left(1 - (0.397)^2\right) = 0.0514 = c.$$

`Anxiety` solely explains 6.1% of the variance of `exam` (with `revision` partialled out), where as `revision` solely explains 1.8%. Therefore `anxiety` is a better predictor of `exam` than `revision`. We can calculate b and e in the following way: $b = R^2 - a - c = 0.209 - 0.0145 - 0.0514 = 0.1431$ and $e = 1 - R^2 = 1 - 0.209 = 0.791$, and then the percentage of variance of Y explained dually by X_1 and X_2 is $b/(b + e) = 0.1431/(0.1431 + 0.791) = 0.1532$.