Overview

Stephanie Ranft

January 14, 2021

Statistics 2
PSBE2-07

Exercises

Regression - ANOVA analysis

1. The "Healthy Breakfast" dataset contains, among other variables, the Consumer Reports ratings of 77 cereals, the number of grams of sugar contained in each serving, and the number of grams of fat contained in each serving.

   Considering "Sugars" as the explanatory variable and "Rating" as the response variable generated the following regression line:

   Rating = 59.3 - 2.40 Sugars

Source	DF	SS	MS	F	p
Regression	1	8654.7	8654.7	102.35	0.000
Error	75	6342.1	84.6		
Total	76	14996.8	194.76		

Table 1: Analysis of Variance - rating ~ sugar

   As a simple linear regression model, we previously considered "Sugars" as the explanatory variable and "Rating" as the response variable.

   The regression line generated by the inclusion of "Sugars" and "Fat" is the following:

   Rating = 61.1 - 2.21 Sugars - 3.07 Fat

Source	DF	SS	MS	F	p
Regression	2	9325.3	4662.6	60.84	0.000
Error	74	5671.5	76.6		
Total	76	14996.8	194.76		
Source	DF	Seq SS			
Sugars	1	8654.7			
Fat	1	670.5			

Table 2: Analysis of Variance - rating ~ sugar + fat

   (a) Define the population regression model using table 2. If two cereals have the same fat content but different sugar content, what can you say about the rating?

   (b) What does VIF stand for? Compute the VIF using table 2.

   (c) What does VAF stand for? Compute the VAF using tables 1 and 2.

   (d) How do the ANOVA results change when "FAT" is added as a second explanatory variable?

   (e) Formulate appropriate hypotheses, make a decision and explain your reasoning.

2. Answer the following questions using the tables and graphs below.

Table 3: Descriptive Statistics

	sales	adverts	airplay	attract
Valid	200	200	200	200
Missing	0	0	0	0
Mean	193.200	614.412	27.500	6.770
Std. Error of Mean	5.706	34.341	0.868	0.099
Std. Deviation	80.699	485.655	12.270	1.395
Minimum	10.000	9.104	0.000	1.000
Maximum	360.000	2271.860	63.000	10.000

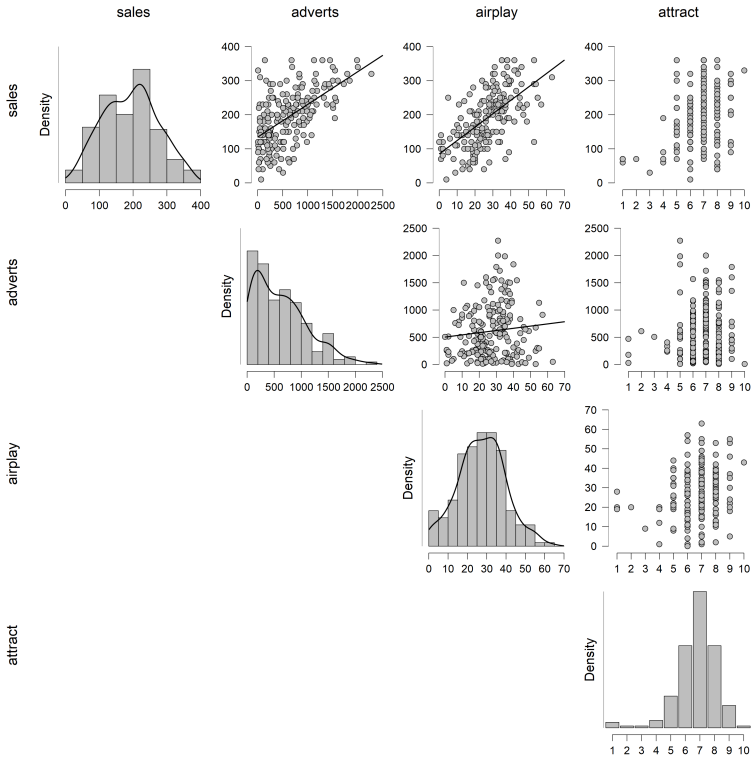


Figure 1

This fictional data set, "Album Sales", provides factors that may influence album sales Variables:

adverts Amount (in thousands of pounds) spent promoting the album before release.

sales Sales (in thousands of copies) of each album in the week after release.

airplay How many times songs from the album were played on a prominent national radio station in the week before release.

attract How attractive people found the band's image (1 to 10).

Table 4: Model Summary

Model	R	R^2	Adjusted R^2	RMSE	R^2 Change	F Change	df1	df2	p
0	0.578	0.335	0.331	65.991	0.335	99.587	1	198	< .001
1	0.815	0.665	0.660	47.087	0.330	96.447	2	196	< .001

Table 5: Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	95% CI		Collinearity Statistics	
							Lower	Upper	Tolerance	VIF
0	(Intercept)	134.140	7.537		17.799	< .001	119.278	149.002		
	adverts	0.096	0.010	0.578	9.979	< .001	0.077	0.115	1.000	1.000
1	(Intercept)	-26.613	17.350		-1.534	0.127	-60.830	7.604		
	adverts	0.085	0.007	0.511	12.261	< .001	0.071	0.099	0.986	1.015
	airplay	3.367	0.278	0.512	12.123	< .001	2.820	3.915	0.959	1.043
	attract	11.086	2.438	0.192	4.548	< .001	6.279	15.894	0.963	1.038

Table 6: ANOVA

Model		Sum of Squares	df	Mean Square	F	p
0	Regression	433687.833	1	433687.833	99.587	< .001
	Residual	862264.167	198	4354.870		
	Total	1.296e+6	199			
1	Regression	861377.418	3	287125.806	129.498	< .001
	Residual	434574.582	196	2217.217		
	Total	1.296e+6	199			

(a) What is the population regression equations for Model 0 and 1?

(b) Describe the regression equations you wrote above, in words (1-2 sentences each model). What is the point of the comparison?

(c) Summarise the findings of table 3 and compare with fig. 1.

(d) Write the null and alternative hypothesis based on the regression equations you wrote in part (a). Now, describe these hypothese in words (do not refer to beta coefficients, just use plain language - like you're informing a friend). What does table 4 inform you about your hypotheses?

(e) Under Model 0, what is the expected output if the explanatory variable input has value 600? Compare this with output with the output from Model 1 under the same conditions. Explain the difference in your results.

(f) Explain the fourth and seventh columns of table 4.

(g) Table 5 provides you with the VIF for both models. Interpret the results without making too many references to the exact value of the VIF, i.e. what do these values mean?

(h) Provide the standardised regression equations for both models.

(i) Use table 6 to make your decision about your hypotheses. Explain your reasoning.

College success: Linear regression and ANOVA

Description:

This data set, "College Success", provides high school grades, SAT scores, and Grade Point Average of 224 university students.

Variables:

`id` Participant ID.

`gpa` Grade Point Average (GPA) after three semesters in college.

`hsm` Average high-school grade in mathematics.

`hss` Average high-school grade in science.

`hse` Average high-school grade in English.

`satm` SAT score for mathematics.

`satv` SAT score for verbal knowledge.

`sex` Gender (labels not available).

We will examine which variables best predict GPA. First, we will fit a model predicting GPA by high school grades. Then, we will use a model that predicts GPA by SAT scores. Finally, we will fit a model that uses both high school grades and SAT scores to predict GPA.

Table 7

(a) Descriptive Statistics

	gpa	hsm	hss	hse	satm	satv
Valid	224	224	224	224	224	224
Missing	0	0	0	0	0	0
Mean	2.635	8.321	8.089	8.094	595.286	504.549
Std. Deviation	0.779	1.639	1.700	1.508	86.401	92.610
Minimum	0.120	2.000	3.000	3.000	300.000	285.000
Maximum	4.000	10.000	10.000	10.000	800.000	760.000

(b) Correlation Table

		gpa	hsm	hss	hse	satm	satv
gpa	Pearson's r	–					
	p-value	–					
	Spearman's rho	–					
	p-value	–					
hsm	Pearson's r	0.436	–				
	p-value	< .001	–				
	Spearman's rho	0.463	–				
	p-value	< .001	–				
hss	Pearson's r	0.329	0.576	–			
	p-value	< .001	< .001	–			
	Spearman's rho	0.395	0.587	–			
	p-value	< .001	< .001	–			
hse	Pearson's r	0.289	0.447	0.579	–		
	p-value	< .001	< .001	< .001	–		
	Spearman's rho	0.316	0.483	0.598	–		
	p-value	< .001	< .001	< .001	–		
satm	Pearson's r	0.252	0.454	0.240	0.108	–	
	p-value	< .001	< .001	< .001	0.106	–	
	Spearman's rho	0.272	0.446	0.237	0.110	–	
	p-value	< .001	< .001	< .001	0.102	–	
satv	Pearson's r	0.114	0.221	0.262	0.244	0.464	–
	p-value	0.087	< .001	< .001	< .001	< .001	–
	Spearman's rho	0.155	0.218	0.271	0.274	0.403	–
	p-value	0.021	0.001	< .001	< .001	< .001	–

What information can be gleamed from table 7a?

1. Do you have to omit any data entries; why/why not?
2. The coefficient of variation (CV) is a standardised measure of dispersion, and often expressed as a percentage. What is the interpretation of c_v = 29.56% for GPA? Compute, interpret, and compare the CV for all six variables. Can you do this for all six variables; why/why not?
3. What does the phrase "no meaningful zero" mean? Explain this in context with regards to one or more of the variables. Does this influence your answer from the previous question?

The correlations in table 7b were produced by JASP using which formula? Why is it important to study correlation in regression? Refer to this table for the following questions.

4. Provide a brief description of the four statistics reported in table 7b.

5. What is the difference between the two correlation statistics reported in the table? When is one more appropriate to use than the other?

6. What can you tell me about the relationship between all six variables (note: you will have to make 15 comparisons[1]).

7. Are there any relationships which you find concerning, and why are you concerned? Provide a reasonable method to solving this problem.

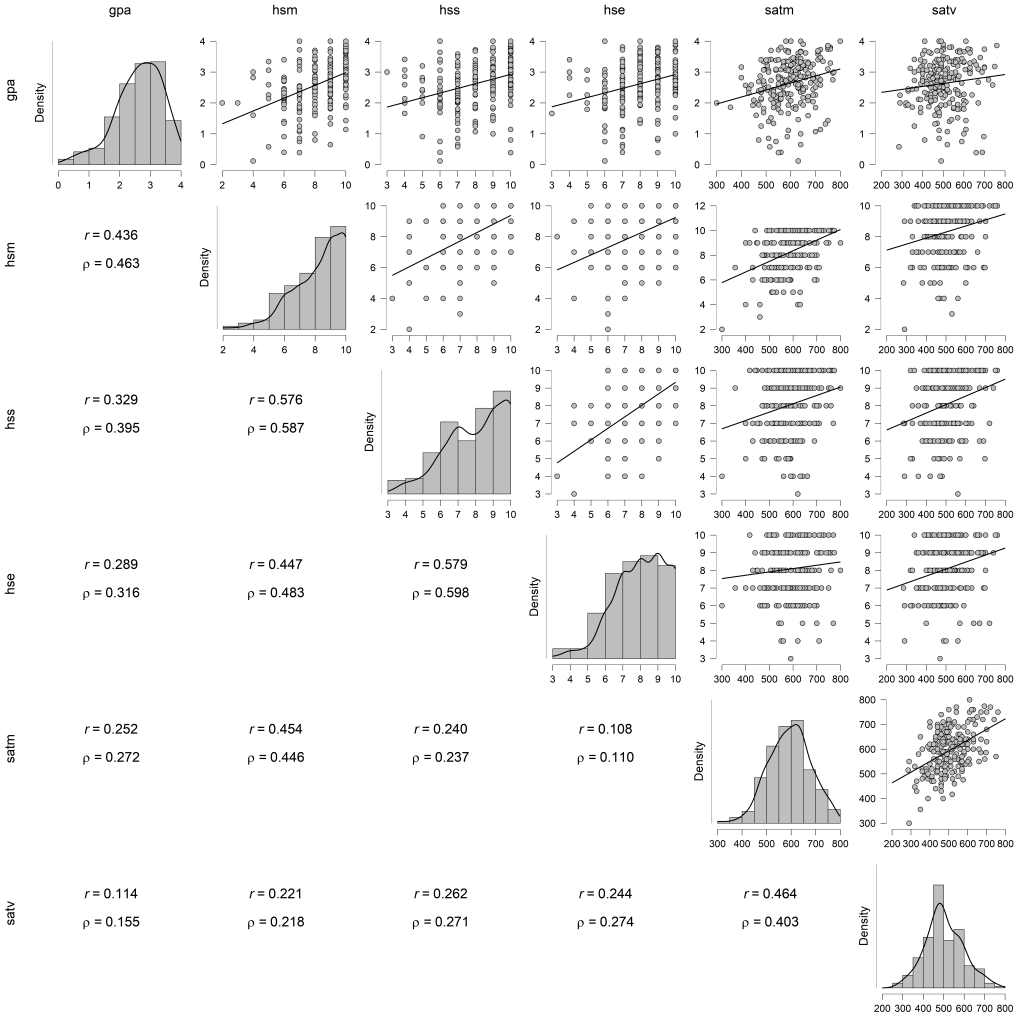


Figure 2

In fig. 2, you are provided with two point estimates and three different graphs in a matrix format, where the lower tri-diagonal contains the point estimates, and the upper tri-diagonal and diagonal contains the graphs.

8. What are the two point estimates and three graph types? Explain the difference.

9. Along the diagonal of the matrix, do you see any graphs which lead to assume there might be violations of linear regression assumptions? If yes, then which graphs might violate which assumptions?

10. With reference to your answer to the previous part, what might be a reason for the observed pattern which leads to a violation? How could you transform your data to solve these problems?

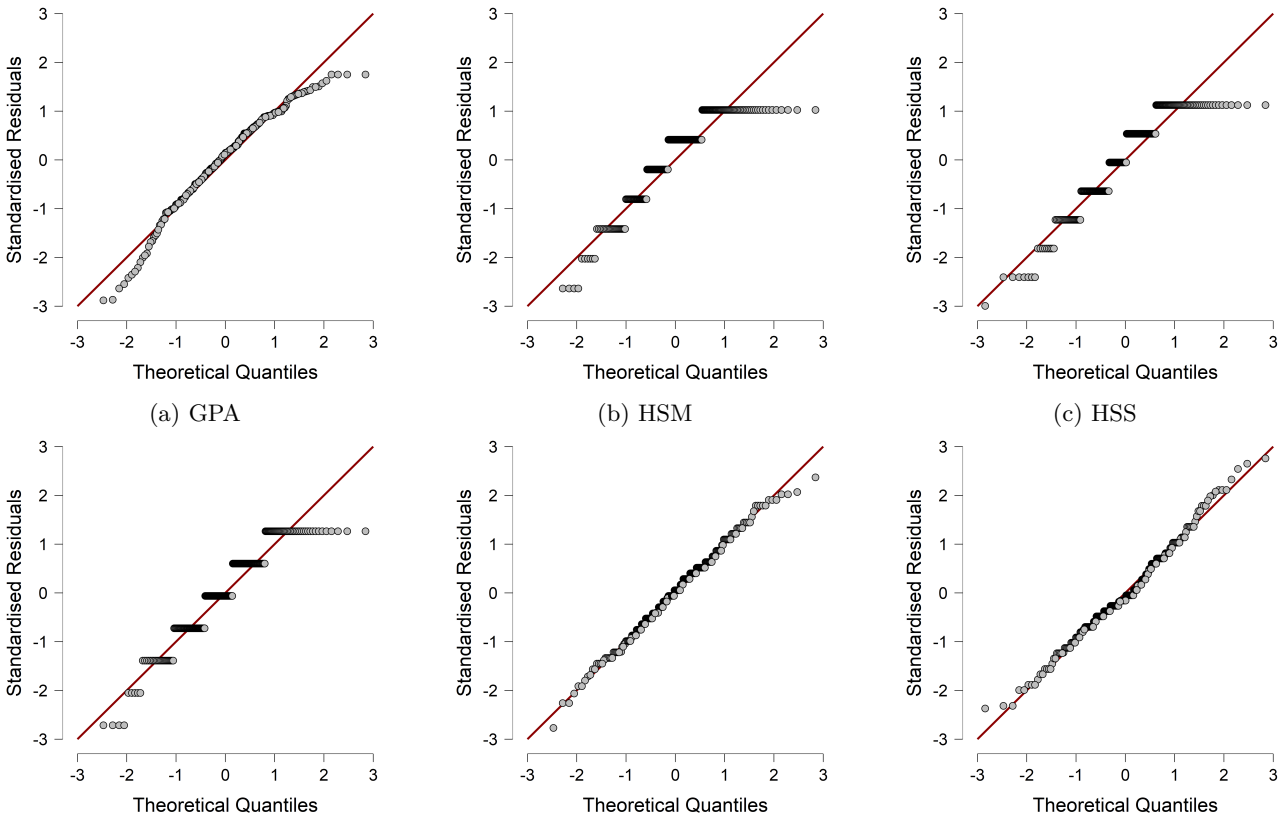
11. Compare the point estimates in fig. 2 to those in table 7b. Do these values agree?

12. Compare the point estimates on the lower tri-diagonal to the graphs on the upper tri-diagonal, and discuss. (Hint: make reference to direction and strength of relationships.)

[1]You have to make 15 comparisons because you have 6 variables and you're going to choose 2 each time to calculate the correlation, i.e. 6 choose 2 = 6!/2!*(6-2)! = 15.

13. Do the graphs on the upper tri-diagonal lead you to believe that there might be a violation of assumption/s? Explain.

14. If you were to standardise the variables, how might the graphs on the upper tri-diagonal change? Why might it be benefit to standardise your variables?



(a) GPA (b) HSM (c) HSS

(d) HSE (e) SATM (f) SATV

Figure 3

Figure 3 displays the Q-Q plots for all six variables. What does "Q-Q" mean? Answer the following questions with reference to this graph.

15. What can you infer about the six variables from fig. 3? How does this compare to your answer given in question 9?

16. What assumption/s are you checking for with a Q-Q plot? Why is this important in inferential frequentist statistics?

17. Write down two equations which represent the assumption/s you presented in the previous question.

Table 8

(a) Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.452	0.205	0.194	0.700

(b) ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	27.712	3	9.237	18.861	< .001
	Residual	107.750	220	0.490		
	Total	135.463	223			

Model		Unstandardized	Standard Error	Standardized	t	p	Collinearity Statistics	
							Tolerance	VIF
1	(Intercept)	0.590	0.294		2.005	0.046		
	hsm	0.169	0.035	0.354	4.749	< .001	0.649	1.540
	hss	0.034	0.038	0.075	0.914	0.362	0.539	1.855
	hse	0.045	0.039	0.087	1.166	0.245	0.645	1.550

(c) Coefficients

Table 8 shows the results of the regression of high-school grades on a dependent variable. What is the population regression model?

18. Explain the four point estimates given in table 8a.

19. Using table 8b, compute the VAF and adjust accordingly. Compare with table 8a.

20. What null hypothesis might you test using table 8b? State the hypothesis/es and make a decision.

21. Explain how the values in each column of table 8b are calculated.

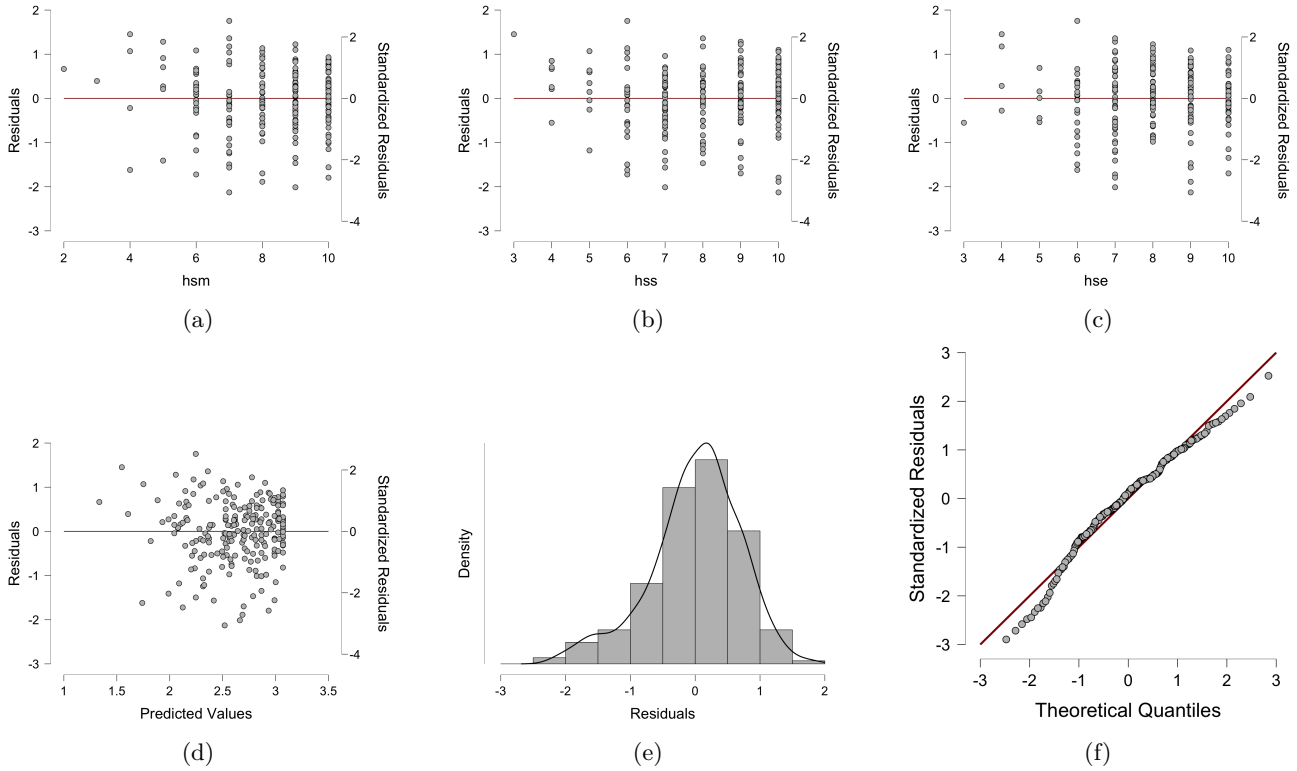
22. Using your population regression model and table 8c, what is the prediction equation? Compute the predicted output for a student with an average high-school mathematics grade of 70%, and explain your answer in words.

23. What can you say about the relevance of the variables in the population model that you have defined? Define hypotheses, test and explain your findings. What might you change/keep the same in the population model and why?

24. What does the VIF column of table 8c tell us? Compare these values with your responses in question 6.

25. How are the columns VIF and Tolerance related? What is meant by the term “tolerance”?

26. In the previous questions, I asked you to investigate individual and joint significance in the population model. Which type responds to which question, and what is the difference?



(a) (b) (c)

(d) (e) (f)

Figure 4

Explain the graphs in fig. 4.

27. You can use figs. 4a to 4c to check which assumption/s? What do you conclude from these graphs?

28. In fig. 4d, the standardised residuals are plotted against the predicted values. Why are the residuals standardised? What can conclude about the assumption of homoskedasticity using this graph?

29. Compare figs. 4e and 4f. How are they different/the same? What assumption are we checking for in this graph? Write the population regression equation which relates to these graphs.

30. With reference to the previous question, are you able to conclude anything about this assumption without any further information?

Now, we include also the SAT scores. Specifically, we include the high school grades in the 'null model'. Then, we add the SAT scores to the model to test whether SAT scores contribute to the prediction of GPA over and above the high-school grades.

Table 9

(a) Model Summary

Model	R	R^2	Adjusted R^2	RMSE	R^2 Change	F Change	df1	df2	p
0	0.452	0.205	0.194	0.700	0.205	18.861	3	220	< .001
1	0.460	0.211	0.193	0.700	0.007	0.950	2	218	0.388

(b) ANOVA

Model		Sum of Squares	df	Mean Square	F	p
0	Regression	27.712	3	9.237	18.861	< .001
	Residual	107.750	220	0.490		
	Total	135.463	223			
1	Regression	28.644	5	5.729	11.691	< .001
	Residual	106.819	218	0.490		
	Total	135.463	223			

Note. Null model includes hsm, hss, hse

Model		Unstandardized	Standard Error	Standardized	t	p	Collinearity Statistics	
							Tolerance	VIF
0	(Intercept)	0.590	0.294		2.005	0.046		
	hsm	0.169	0.035	0.354	4.749	< .001	0.649	1.540
	hss	0.034	0.038	0.075	0.914	0.362	0.539	1.855
	hse	0.045	0.039	0.087	1.166	0.245	0.645	1.550
1	(Intercept)	0.327	0.400		0.817	0.415		
	hsm	0.146	0.039	0.307	3.718	< .001	0.531	1.884
	hss	0.036	0.038	0.078	0.950	0.343	0.532	1.878
	hse	0.055	0.040	0.107	1.397	0.164	0.617	1.620
	satm	9.436e-4	6.857e-4	0.105	1.376	0.170	0.626	1.597
	satv	-4.078e-4	5.919e-4	-0.048	-0.689	0.492	0.731	1.367

(c) Coefficients

Answer the following questions using table 9.

31. What is the population regression equation for Model 1, and what is it estimated to be?

32. Explain table 9a. Making reference to the equation you defined in the previous question, what is your null and alternative hypotheses? What kind of test do you perform?

33. Explain why some values in table 9b are similar/the same, and explain why some values are different. Why does the value of F change between the models; what does this tell us?

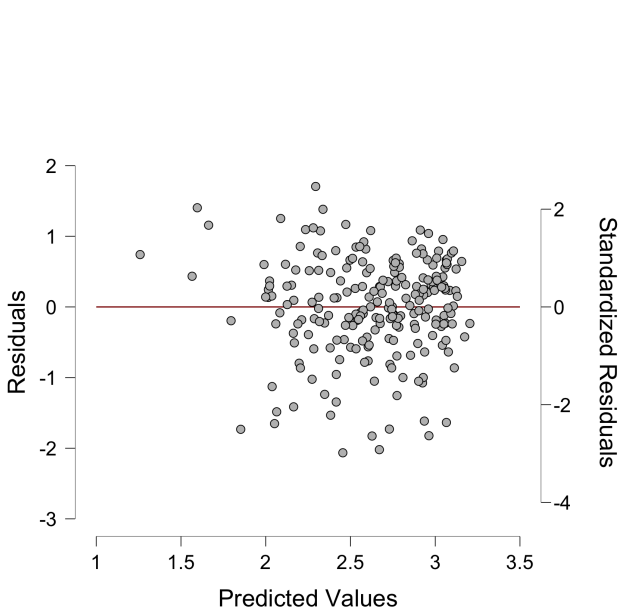
34. The estimated beta coefficients are given in table 9c. Explain the difference in the intercept and slope values.

35. What do you conclude about the individual significance of the beta coefficients in Model 1? How many tests are required for this comparison?

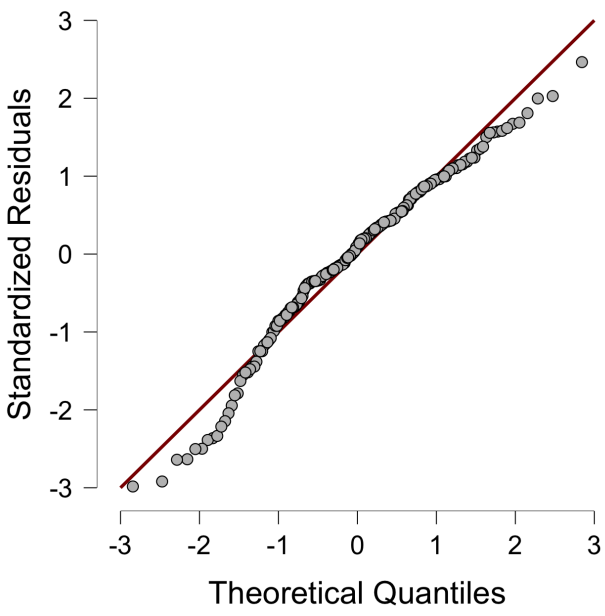
36. One of the standardised beta coefficients is negative. Why is that?

37. Are there any violations of multicollinearity present in either model? Explain.

38. Is the addition of SAT scores to the population model warranted? Why/ why not?



(a)



(b)

Figure 5

Write 4-5 lines describing why we need to examine figs. 6a and 6b and what you conclude (4-5 lines for each graph).

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 10

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:

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

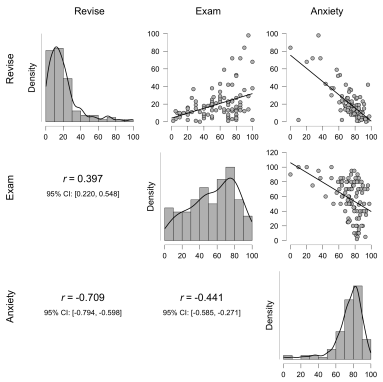


Figure 7

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. 7, 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. 7, 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.

$$pr^2_1 = \frac{R^2 - r^2_{Y,X_2}}{1 - r^2_{Y,X_2}}; \qquad pr^2_2 = \frac{R^2 - r^2_{Y,X_1}}{1 - r^2_{Y,X_1}};$$
$$sr^2_1 = pr^2_1 \cdot \left(1 - r^2_{Y,X_2}\right) = R^2 - r^2_{Y,X_2}; \qquad sr^2_2 = pr^2_2 \cdot \left(1 - r^2_{Y,X_1}\right) = R^2 - r^2_{Y,X_1}.$$

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.

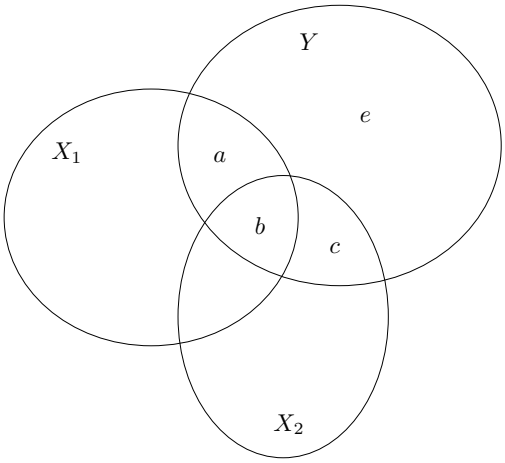


Figure 8