Formulas

Stephanie Ranft S2459825

October 31, 2019

Statistics 2
PSBE2-07

Exercises

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 1

(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 1a?

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 1b 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 1b.
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.

[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.

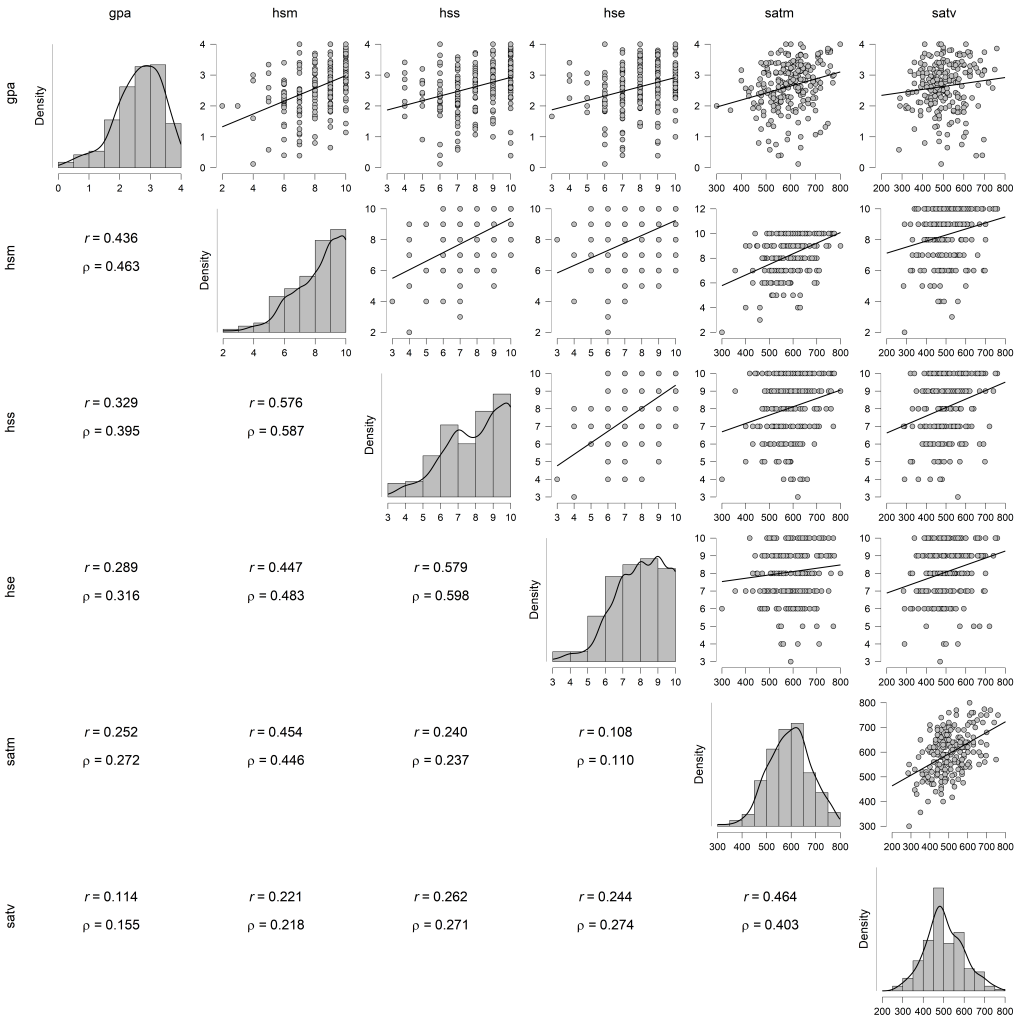


Figure 1

In fig. 1, 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?

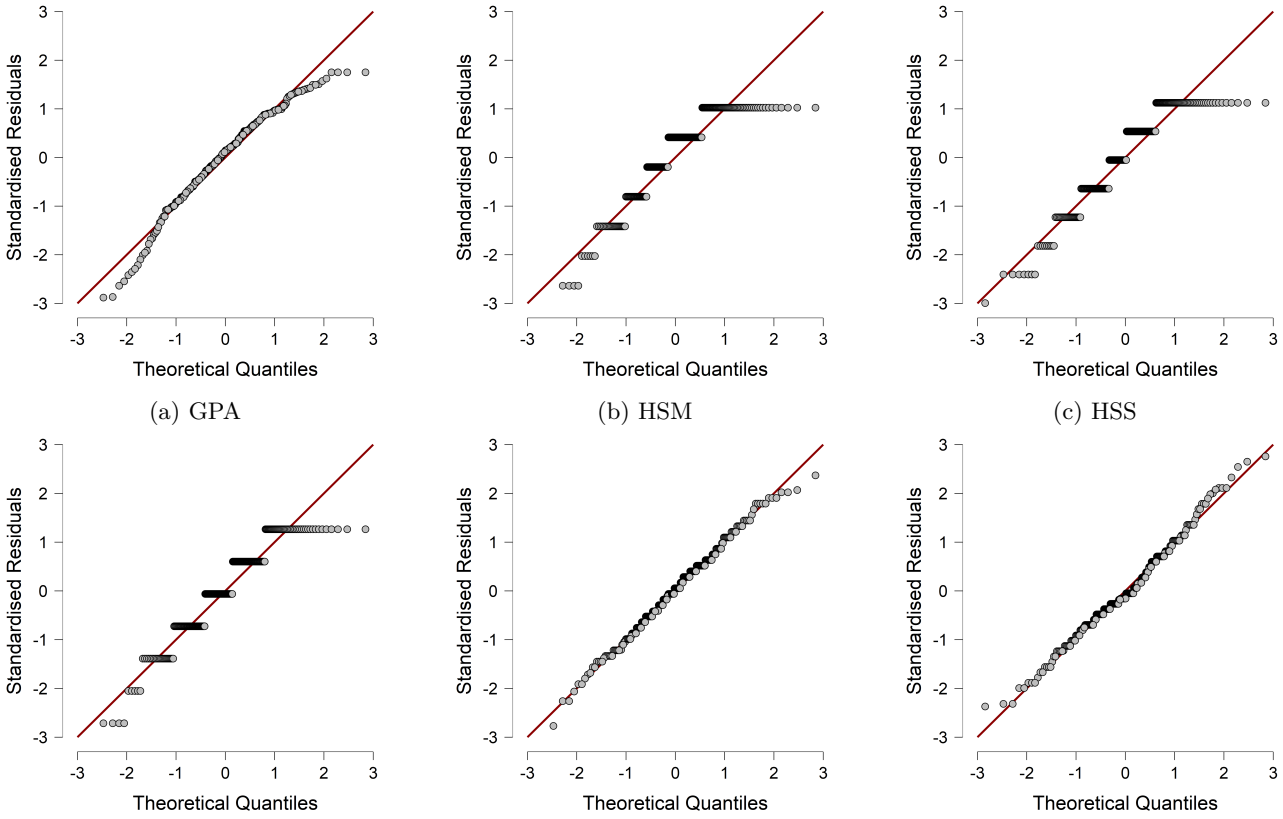
10. With reference to your answer to the previous part, what might be a reason for the observed pattern which leads to a violation? How could you transform your data to solve these problems?

11. Compare the point estimates in fig. 1 to those in table 1b. 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.)

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 2

Figure 2 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. 2? 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 2

(a) Model Summary

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

(b) ANOVA

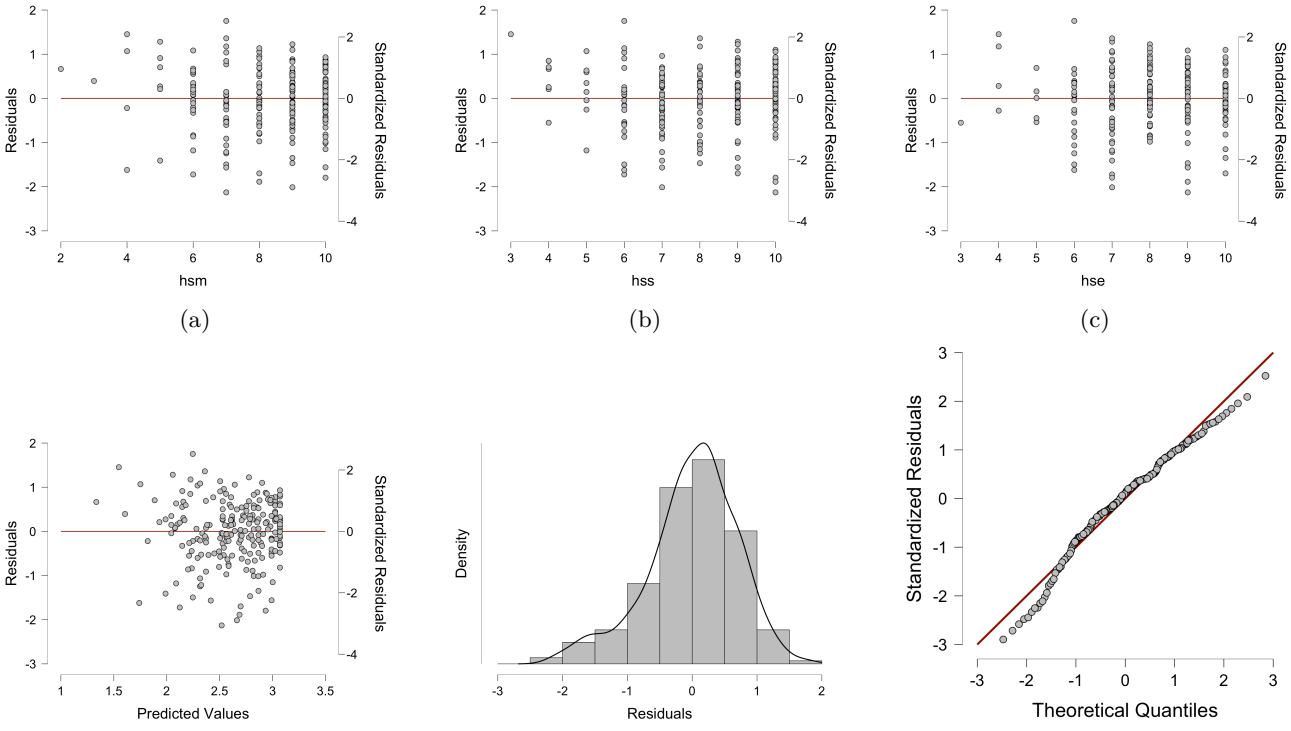
Model		Sum of Squares	df	Mean Square	F	p
1	Regression	27.712	3	9.237	18.861	< .001
	Residual	107.750	220	0.490		
	Total	135.463	223			

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 2 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 2a.
19. Using table 2b, compute the VAF and adjust accordingly. Compare with table 2a.
20. What null hypothesis might you test using table 2b? State the hypothesis/es and make a decision.
21. Explain how the values in each column of table 2b are calculated.
22. Using your population regression model and table 2c, 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 2c 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 3

Explain the graphs in fig. 3.

27. You can use figs. 3a to 3c to check which assumption/s? What do you conclude from these graphs?

28. In fig. 3d, 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. 3e and 3f. 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 3

(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			

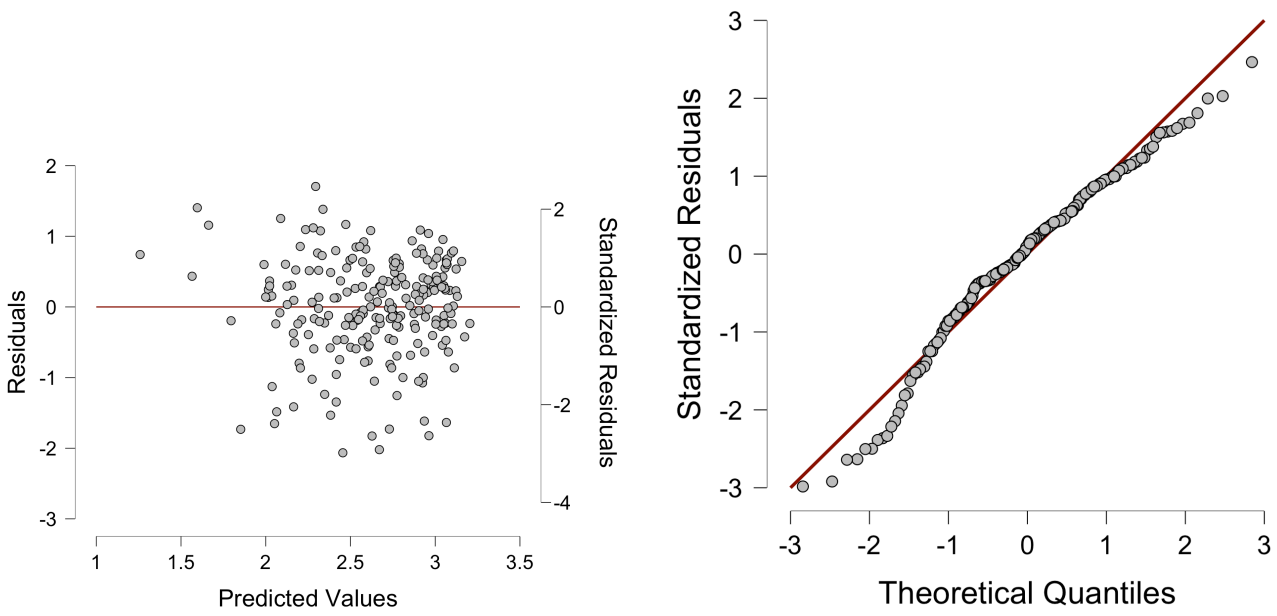
Note. Null model includes hsm, hss, hse

							Collinearity Statistics	
Model		Unstandardized	Standard Error	Standardized	t	p	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 3.

31. What is the population regression equation for Model 1, and what is it estimated to be?
32. Explain table 3a. 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 3b 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 3c. 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 4

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