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

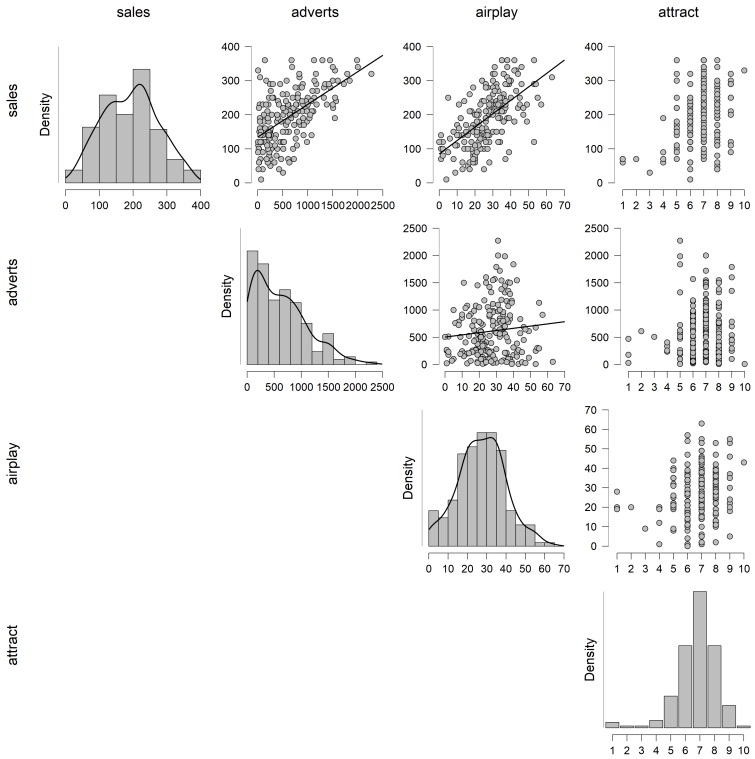


Figure 1

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.

Solutions

1. (a) The population regression model used in table 2 is rating = β_0 + β_1 sugars + β_2 fat + ε, where β_j's are approximated by b_j's such that $\vec{b} = (61.1, -2.21, -3.07)^T$. If variable fat is kept constant, then the marginal difference in rating is -2.21 per gram of sugar. This says that the rating of the breakfast cereal will decrease by 2.21 points per additional gram of sugar, under the condition that fat content is kept constant.

   (b) VIF stands for variance inflation factor, and is given by $\text{VIF}_j = 1/\left(1 - R_j^2\right)$, where R_j^2 is the coefficient of determination of the regression equation $X_j = \alpha_0 + \alpha_1 X_{-j} + \delta$ (regress the explanatory variables on the others). The square root of the VIF indicates how much larger the standard error increases compared to if that variable had 0 correlation to other predictor variables in the model. For example, if the variance inflation factor of a predictor variable were 5.27 ($\sqrt{5.27} = 2.3$), this means that the standard error for the coefficient of that predictor variable is 2.3 times larger than if that predictor variable had 0 correlation with the other predictor variables. Rule of thumb: $VIF_j > 10$ indicates multicollinearity in the model, i.e. explanatory variables are dependent on each other.

   It is not possible to compute the VIF using table 2, as we need the partial SS information.

   (c) VAF stands from variance accounted for and is given by the R^2 coefficient for linear regression, where $R_2 = SSR/SST$. From table 1, R^2 = 8654.7/14996.8 = 0.577, and from table 2, R^2 = 9325.3/14996.8 = 0.622.

   (d) Comparing the VAF's tells us that the model is improved with the addition of fat to the model. This is further shown by the column Seq SS, which shows that fat reduces the SSE by 670.5, which in turn reduces the MSE, indicating less deviation between the observed and fitted values.

   (e) $H_0 : \beta_2 = 0$ and $H_A : \beta_2 \neq 0$. F is significant with $p < 0.05$, i.e. reject H_0 in favour of H_A and conclude that fat is in the population model.

2. See the attached JASP file.

   (a) Model 0: sales = β_0 + β_1 adverts + ε.
   Model 1: sales = β_0 + β_1 adverts + β_2 airplay + β_3 attract + ε.

   (b) Model 0: The sales (in thousands of copies) of each album in the week after release depends only on the amount (in thousands of pounds) spent promoting the album before release.

   Model 1: The sales (in thousands of copies) of each album in the week after release depends on the amount (in thousands of pounds) spent promoting the album before release, how many times songs from the album were played on a prominent national radio station in the week before release, and how attractive people found the band's image (on a scale from 1 to 10).

   The point is to see whether album sales depend only on how much is spent on advertising, or if there is a "organic" component to the music industry, and that album sales still depend on radio airtime and fans. It is clear to all that the music industry has progressed from this organic state to a hyper-commercialised money-making machine, and we wish to test if this formula prescribed by the big music producers is actually what drives sales (i.e. money in their pockets), or if people still care about the artists as people and that the music is subjectively good.

   (c) Table 3 gives us the descriptive statistics of all four variables. We can use this information to create confidence intervals for the population means (of each variable). Figure 1 shows the correlations between all four variables, as well as the distribution of each variable. It is evident from the first row that sales is correlated with adverts, airplay and attract, and that the latter three are uncorrelated with each other. The row for adverts shows that adverts is slightly right skewed, however we can be sure of any violations of normality by viewing the QQ-plots - we do not have enough information from the density plots to make a decision.

   (d) $H_0 : \beta_2 = 0$ and $\beta_3 = 0$.
   $H_1 : \beta_2 \neq 0$ and $\beta_3 \neq 0$.

   Note: these hypotheses talk about whether airplay and attract are **jointly significant** in the model, meaning that we must conduct an F-test. If we reject H_0, we do not know anything about the values of β_2 and β_3 in the population model, **only that they are both not zero**. It could be that $\beta_2 = 0$ and $\beta_3 \neq 0$ (attract is **individually significant** in the model, but airplay isn't), or vice versa, or maybe even that they are both **individually significant** in the model i.e. both coefficients not significantly zero in the model. In order to ascertain which of the two (or both) are non-zero, we would need to perform multiple t-tests however this presents an issue with a too large Type-I error (Bonferroni). This is why we perform an F-test first, to see whether it is worth our time to perform any t-tests.

If we find sufficient evidence in the data, we can conclude that album sales depends on the amount of money spent on advertising, the radio airtime, and the band's public image. However if there isn't sufficient evidence, then we must conclude that the album sales solely depend on the amount of money spent on advertising.

table 4 tells us that both Model 0 and 1 are "good" as the F-test shows (see the p-value), however Model 1 is "better" as shown by the R or R^2 columns. Furthermore, the adjusted R^2 column shows that there is no issue of "over-fitting", i.e. we haven't included too many predictors in the model. The R^2 change column tells us that adding airplay and attract to the model increases the VAF by 33% - the F change column is an F-test done on the VAF, which shows significant difference. All of this just says that Model 0 is ok, but Model 1 is better.

(e) Under Model 0,

$$\widehat{\text{sales}}\,(\text{adverts} = 600) = 134.140 + 0.096 \times 600 = 191.74,$$

or equivalently "spending 6,000 pounds on advertising yields an estimated 191,740 in album sales". Under Model 1,

$$\begin{aligned}\widehat{\text{sales}}\,(\text{adverts} = 600) &= -26.613 + 0.085 \times 600 + 3.367\text{airplay} + 11.086\text{attract} \\ &= 24.387 + 3.367\text{airplay} + 11.086\text{attract},\end{aligned}$$

or equivalently "spending 6,000 pounds on advertising yields an estimated 24,387 in album sales, keeping other factors constant". The value of the coefficient for adverts in either model are similar which shows that the marginal contributions are the same (confirm this with the standardised coefficient, i.e. the r between sales and adverts), however the intercepts are very different so keeping other factors constant does not provide an accurate estimate in Model 1. In Model 1, airplay and attract capture some of the variation in sales and so the marginal contribution of adverts to sales is smaller in Model 1 than in Model 0.

(f) See part (d).

(g) The VIF tells us whether adverts, airplay and attract are correlated with each other. If, for example, adverts and attract were correlated, then the R_j^2 produced from the following regression

attract $= \alpha_0 + \alpha_1$ adverts $+ \delta$

would be closer to 1 than to zero. Then, instead of including attract in the model, we would reformulate Model 1 as follows:

$$\begin{aligned}\text{sales} &= \beta_0 + \beta_1 \text{ adverts} + \beta_2 \text{ airplay} + \beta_3 (\alpha_0 + \alpha_1 \text{ adverts} + \delta) + \varepsilon \\ &= (\beta_0 + \beta_3\alpha_0) + (\beta_1 + \beta_3\alpha_1) \text{ adverts} + \beta_2 \text{ airplay} + (\varepsilon + \beta_3\delta).\end{aligned}$$

We have a new intercept term $(\beta_0 + \beta_3\alpha_0)$, a new slope term for adverts $(\beta_1 + \beta_3\alpha_1)$, and a new error term $(\varepsilon + \beta_3\delta)$. The new intercept and slope are interesting however the new error term is not interesting for us - we can rename it as $\tilde{\varepsilon}$.

sales $= (\beta_0 + \beta_3\alpha_0) + (\beta_1 + \beta_3\alpha_1)$ adverts $+ \beta_2$ airplay $+ \tilde{\varepsilon}$.

VIF is calculated as $1/(1 - R_j^2)$ and the rule of thumb is that if VIF>10, then there is multicollinearity between X_j and the other predictors - a VIF>10 corresponds with an $R_j^2 > 0.9$, i.e. strong correlation. The VIF columnn in table 5 tells us that the there is no issue of multicollinearity.

(h) Model 0: sales = 0.578×adverts.

Model 1: sales = 0.511×adverts + 0.512×airplay + 0.192×attract.

(i) Table 6 tells us that the MSE is smaller in Model 1, when comparing it to the MSE in Model 0, which tells us that the addition of the predictors airplay and attract to the model has had a positive effect on the reliability of future predictions of album sales based on the data. The p-values say that the MSE is low enough for both models for us to conclude that linear regression is appropriate, however the MSE cells tell us that Model 1 is better.

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.

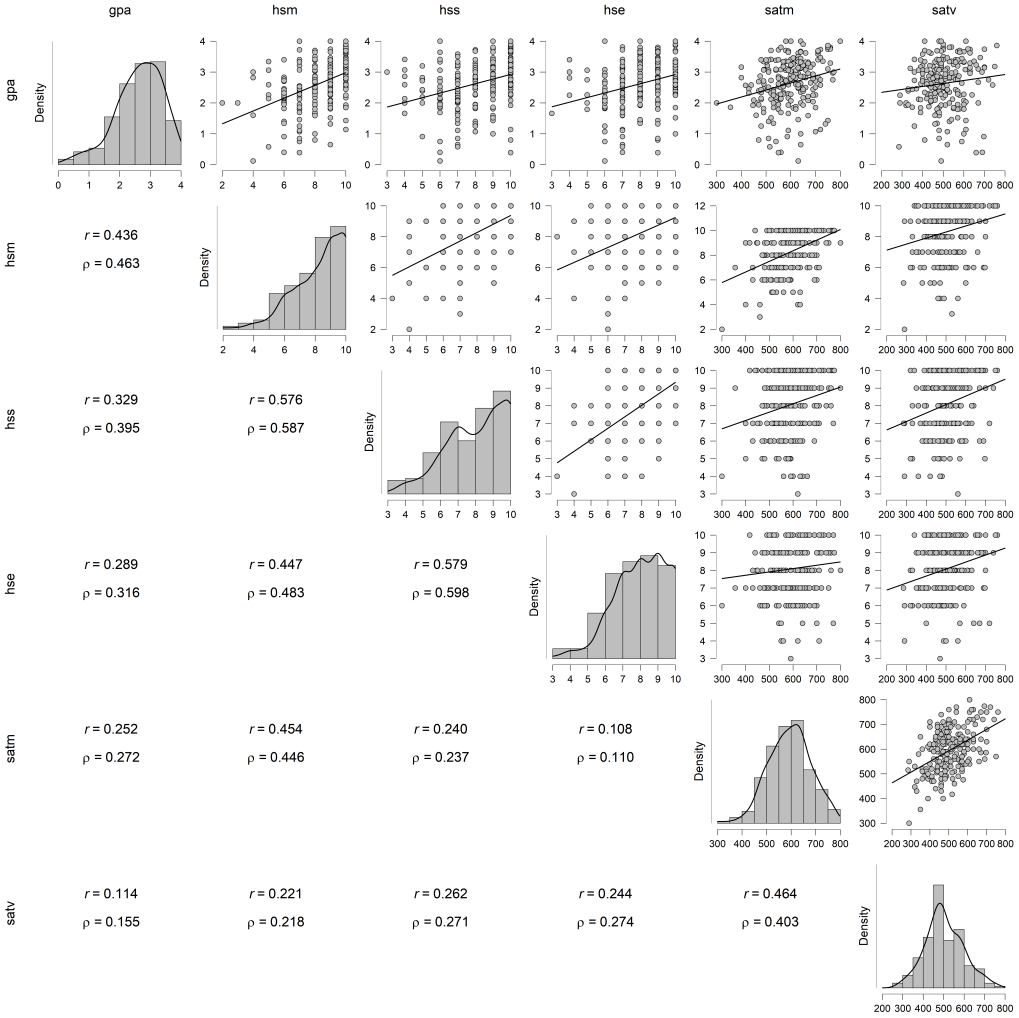


Figure 2

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?

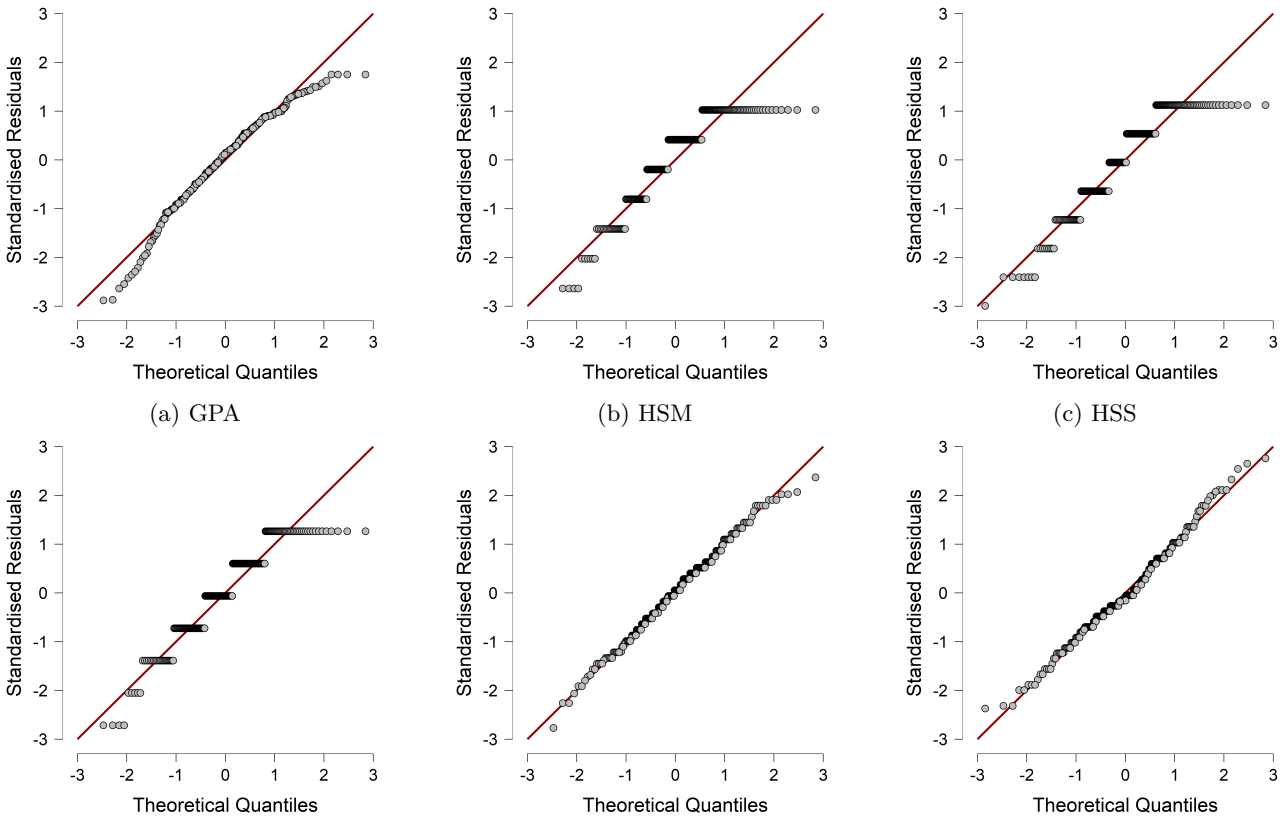
11. Compare the point estimates in fig. 2 to those in table 7b. Do these values agree?

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.

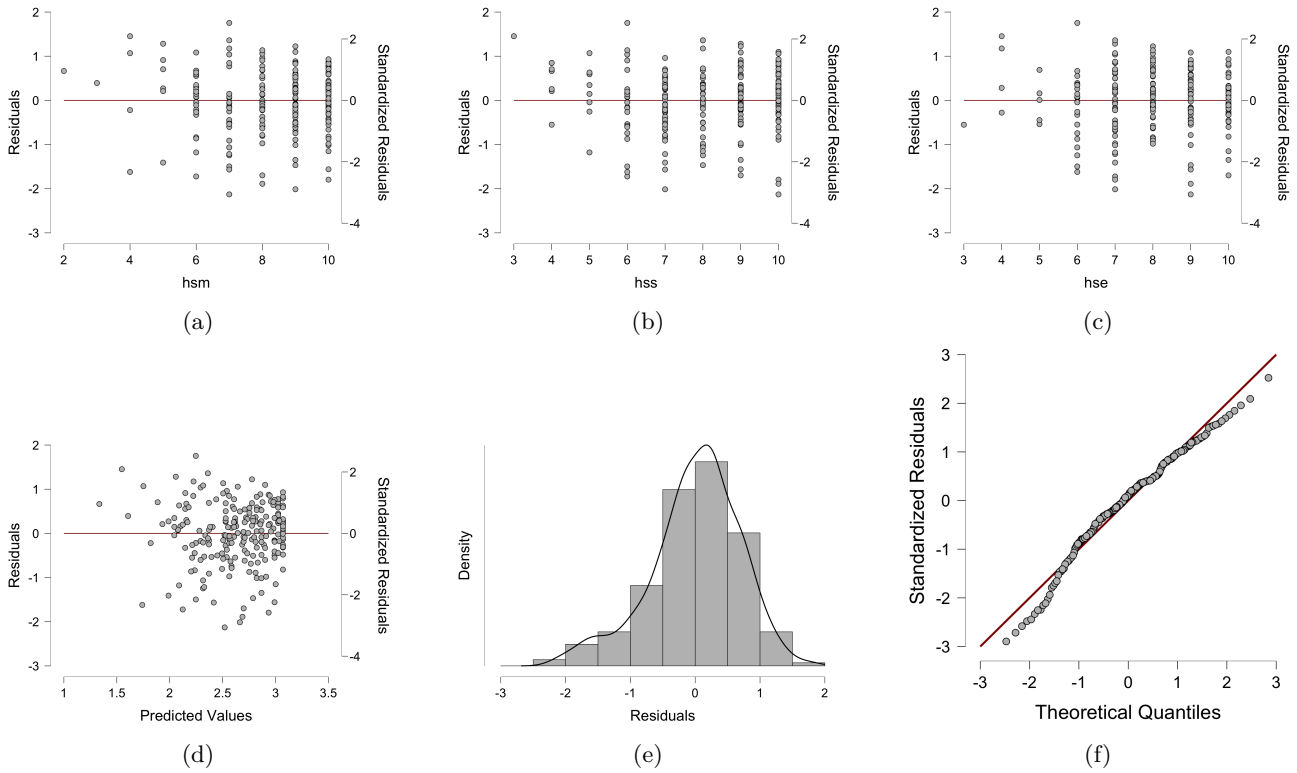
22. Using your population regression model and table 8c, what is the prediction equation? Compute the predicted output for a student with an average high-school mathematics grade of 70%, and explain your answer in words.

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?

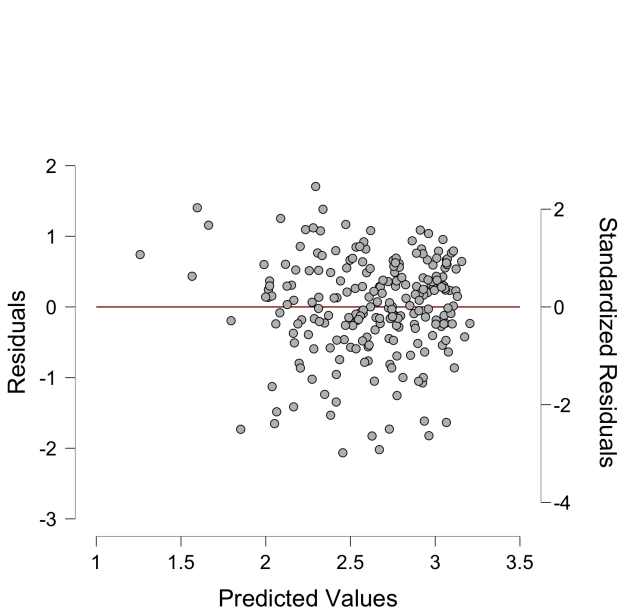
34. The estimated beta coefficients are given in table 9c. Explain the difference in the intercept and slope values.

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

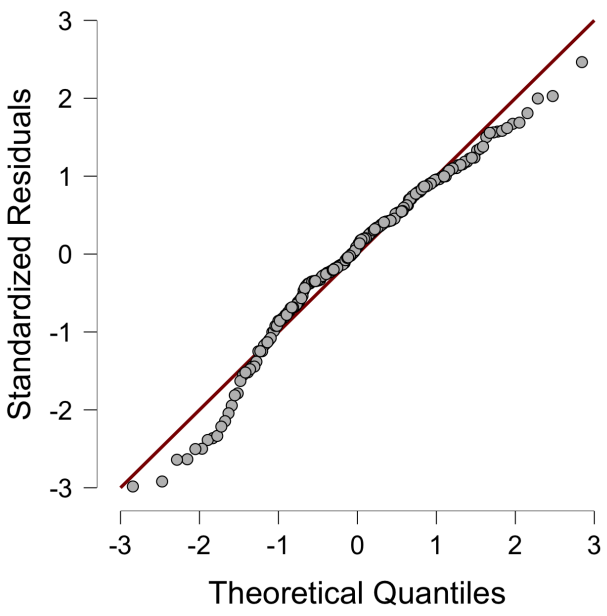
36. One of the standardised beta coefficients is negative. Why is that?

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).

Solutions

1. No; all data points are valid.

2. CV shows the extent of variability in relation to the mean of the population. You do not interpret a CV alone, rather use it to compare with other variables.

$$c_v^{HSM} = \frac{1.639}{8.321} = 19.70\%; \qquad c_v^{HSS} = \frac{1.700}{8.089} = 21.02\%; \qquad c_v^{HSE} = \frac{1.508}{8.094} = 18.63\%;$$
$$c_v^{SATM} = \frac{86.401}{595.286} = 14.51\%; \qquad c_v^{SATV} = \frac{92.610}{504.549} = 18.36\%.$$

   The variables `hss`, `hse`, and `hsm` are expressed on an integer scale from 1 to 10 (interval scale), and `gpa`, `satm`, and `satv` are expressed on different ratio scales. Therefore it only makes sense to compare the c_v's of `gpa`, `satm`, and `satv`. Clearly, `gpa` has more variability than `satv`, and nearly twice as much as `satm`.

3. This phrase means that there is a value attainable by the variable which indicates the state of nothing. In this case, `satm` and `satv` have a meaningful zero point which indicates receiving no points on the exams, and `gpa` has a meaningful zero point indicating a grade average of zero. The variables `hss`, `hse`, and `hsm` do not have a meaningful zero point, as the grades start at 1.

4. Pearson's r gives the linear correlation coefficient which displays direction and strength of the relationship between two variables. The associated p-value gives the probability of observing r under the assumption that the two variables are independent. If the p-value is less than our prescribed error minimum, then we must accept that the variables are dependent.

   Spearman's ρ is the non-parametric version of Pearson's r which relies solely on the rank of paired data, i.e. if most pairs increase together then $\rho \approx 1$, however if most pairs move in opposite directions to each other then $\rho \approx -1$. However, if the data pairs do not have a majority moving together or apart, and the "movement" is random, then $\rho \approx 0$. The association p-value has the same interpretation.

5. See above. Pearson's for quantative data and Spearman's for qualitative data.

6. `gpa` is significantly (at the 5% level) correlated with all variables (look at Pearson's for `satm` and `satv` and Spearman's for `hsm`, `hss`, and `hse`), except `satv` with a p-value of ~ 8% and therefore is significant at the 10% level. `hsm` and `hss` are significantly correlated with all other variables, and `hse` fails to be significantly correlated with `satm` at the 10% level.

7. We are concerned about multicollinearity in the predictors, so it may be better to regress `hse` on `hsm` and `hss`, and then regress `satm` on `satv` so that the model is now `gpa` = β_0 + β_1 `hse` + β_2 `satm` + ε. This will also provide a partial solution to the low correlation present between `gpa` and `satv`.

8. Pearson's r and Spearman's ρ (see above). Along the diagonal are the distributions of the variables. On the upper tri-diagonal are the scatter plots between the two variables. It is clear to see that `hsm`, `hss` and `hse` have interval scale type graphs, and `gpa`, `satm` and `satv` have ratio scale type graphs. The interval scale graphs show the data points in groups of straight lines, whereas the ratio scale type graphs show a cloud of data points.

9. `gpa` seems slightly skewed to the left, however `hsm`, `hss` and `hse` are greatly skewed to the left. This seems appropriate as the grades are bounded above and grade averages tend toward the upper bound (students try to get top grades). The graphs of `satm` and `satv` seem relatively normal, however `satv` appears to have slightly positive kurtosis (sharp peak). The normality of these two graphs makes sense as these are the scores of a single test, rather than an average of grades, and therefore an even distribution is expected. The slight positive kurtosis might be best explained by the graph itself: the bar indicating the mode in the histogram is almost double in length than the other proceeding bars, i.e. the number of people who got the median score is about double the number of people for any other score. This might be due to most people having a below-average level of reading and writing (in the US).

10. See above. Instead of performing tests on means and standard deviations, you could test on medians and IQR's, or use Bayesian methods. You could try to normalise the skewed variables and ease the kurtosis with JASP or SPSS.

11. Yes.

12. All positively correlated. Steep slope indicates strong linear correlation, and flat slope indicates weak linear correlation.

13. We want the slopes of the graphs on the top line to be steep, and the graphs on the lines below to be quite flat. This is not what is observed so there may not be a strong linear relationship between the dependent and independent variables, and there might a violation of the multicollinearity assumption.

14. Standardising the variables would result in the slope of the graph being exactly Pearson's r, which allows a much quicker picture of the linear relationship than the unstandardised variables. This may also resolve slight skewing and kurtosis on the diagonal (not much, but a little).

15. Q-Q plots compare the distributions of two variables using theoretical quantiles. If the two variables have the same distribution, then the data points line up along the $y = x$ line. By assumption, the standardised residuals have the standard normal distribution, so we are determining if the variables are roughly normally distributed. We confirm our suspicions about the skewing but conclude that it is not terrible (difference is about 1). Similarly, we confirm that `satm` is normal and the kurtosis in `satv` is negligible.

16. Checking for normality, as we use the Central Limit Theorem to provide inferences about the population (assumed normal) using sample data.

17. We assume that the dependent variable is normally distributed in the population (CLT) with some mean μ_Y and some variance σ_Y^2, and that the data is sampled in a way which ensures independence between the Y_i's. We assume that the independent variables are uncorrelated with each other (multicollinearity) and uncorrelated with the error term. We assume that the the error term is normally distributed with mean zero (we expect no errors) and with the same variance as Y (variance of error does not depend on independent variables: homoskedasticity), and that the error terms are not correlated with each other.

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}_{\mu_Y} + \varepsilon; \qquad Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right); \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma_Y^2\right).$$

18. R is actually Pearson's r for the regression equation, i.e. combined linear correlation coefficient. R^2 is the variance accounted for (VAF) by the independent variables (of the dependent variable). Adjusted R^2 is the VAF adjusted for the number of predictors; adding predictors to the model increases R^2 whether there is any true benefit to the MSE, so R^2 is adjusted downwards relative to the number of predictors added (Wherry's). RMSE (root mean squared error) is the standard error for the estimate of R.

19. $R^2 = SSR/SST = 27.712/135.463 = 0.205$. $\bar{R}^2 = 1 - \left(1 - R^2\right) \cdot (n-1) \,/\, (n-p-1) = 1 - (1 - 0.205) \cdot (223)/(223 - 4) = 0.194$.

20. $\texttt{gpa} = \beta_0 + \beta_1\ \texttt{hsm} + \beta_2\ \texttt{hss} + \beta_3\ \texttt{hse} + \varepsilon$.

    $H_0 : \beta_1 = 0$ and $\beta_2 = 0$ and $\beta_3 = 0$, i.e. `hsm`, `hss`, and `hse` are jointly insignificant in the model.

    $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$, i.e. `hsm`, `hss`, and `hse` are jointly significant in the model.

    Notice the use of "and" and "or" in the hypotheses - very important. The null says that all beta coefficients are zero in the population model, i.e. dependent is uncorrelated with the independent variables. The alternative says that at least one of the beta coefficients is non-zero, i.e. at least one of the predictors is correlated with the dependent variable.

    The value of $F(3, 220) = 18.861$ is significant $(p < 0.001)$ therefore we reject the null hypothesis.

21. SS column is additive: SST = SSR + SSE. df column is additive: dfT $= n-1 =$ dfR + dfE $= (p-1)+(n-p)$. MSR = SSR / dfR and MSE = SSE / dfE, and then F = MSR / MSE. $p = \mathbb{P}\left(F_{3,220} > 18.861\right)$ found using software or an F-table.

22. $\widehat{\texttt{gpa}} = 0.59 + 0.169\ \texttt{hsm} + 0.034\ \texttt{hss} + 0.045\ \texttt{hse}$. If `hsm` = 7, then we can input this into the prediction equation and set `hss` and `hse` to meaningful zeros, i.e. to their respective means: $\widehat{\texttt{gpa}} = 0.59 + 0.169 * 7 + 0.034 * 8.089 + 0.045 * 8.089 = 2.412$. We can interpret this as, on average a student with a high-school math grade of 7 out of 10 has a GPA in university of 2.412.

23. Look at the p-values: only `hsm` is individually significant.

24. Rule of thumb (SL 17; L06): VIF < 4 is good. Previously we were concerned about multicollinearity however all VIF's are below 2.

25. VIF = 1 / Tolerance, where Tolerance = 1 - R_j^2 is percentage of variance of X_j which is not explained by the other predictors.

26. Individual is t-tests on single beta coefficients. Joint is an F-test on all beta coefficients (or a range).

27. Homoskedasticity and uncorrelated error terms. No violations as the pattern of the data points is random and bounded by parallel lines.

28. The residuals are standardised so you can get a visual estimate of the variance of the residuals. The left axis gives the unstandardised residual values. Homoskedasticity assumption not violated as the pattern of the data points is random and bounded by parallel lines.

29. See the answer to question 17.

30. Yes, we conclude that the assumptions of normality (for both the dependent variable and the error term) and equal variance are not violated.

31.

$$\texttt{gpa} = \beta_0 + \beta_1\texttt{hsm} + \beta_2\texttt{hss} + \beta_3\texttt{hse} + \beta_4\texttt{satm} + \beta_5\texttt{satv} + \varepsilon$$
$$\widehat{\texttt{gpa}} = 0.327 + 0.146\texttt{hsm} + 0.036\texttt{hss} + 0.055\texttt{hse} + 0.0009346\texttt{satm} + 0.0005919\texttt{satv}$$

32. $H_0 : \beta_4 = 0$ and $\beta_5 = 0$. $H_1 : \beta_4 \neq 0$ or $\beta_5 \neq 0$. This is an F-test for joint significance. We look at the F change column and note that we do not reject our null hypothesis.

33. The SSR's are quite similar, however the dfR's are different and therefore result in different MSR's. As the SSR's are similar (and for obvious reasons the SST's are the same), the SSE's are similar and despite the difference in dfE's, the MSE's are equal. The different MSR's and equal MSE's result in different F statistics. This says that the addition of `satv` and **satm** to the model does not decrease the MSE significantly, which agrees with the answer given in the previous question.

34. The slope values do not vary greatly which suggests that the predictors in Model 0 are better than those which were included in Model 1. The intercept values change by approx. 0.2 as the inclusion of the variables has altered the slope coefficients slightly, moreover the model now has variables with zero-valued meaningful zeros (whereas Model 0's meaningful zeros are the means of the variables).

35. We require 5 t-tests for individual significance, and it is clear that only `hsm` is individually significant.

36. The linear relationship between `gpa` and `satv` is negative, i.e. those with a higher SAT verbal score usually do worse in hard-science subjects, such as maths, and high maths grades is a stronger positive contributor to university GPA. The negative correlation coefficient is quite near zero, and table 7b stated insignificant correlation between `gpa` and `satv`.

37. No. All VIF < 2.

38. No. Reject null hypothesis.

See the answer to question 17.

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:

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

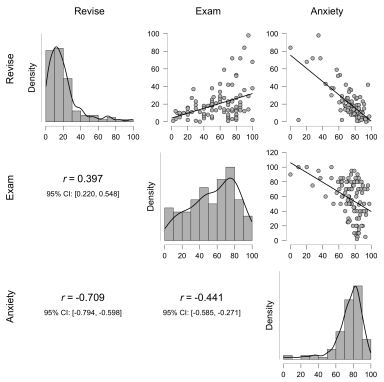


Figure 7

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.

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

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.

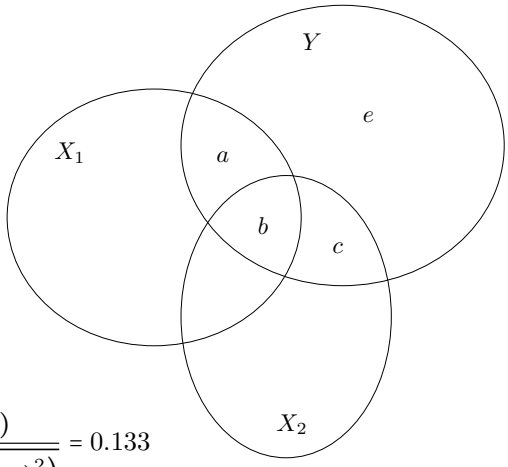


Figure 8

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 $b/(b+e) = 0.1431/(0.1431 + 0.791) = 0.1532$.