Formulas

Stephanie Ranft S2459825

October 17, 2019

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.

   Solution.

   (a) The population regression model used in table 2 is rating = $\beta_0 + \beta_1$ 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. 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

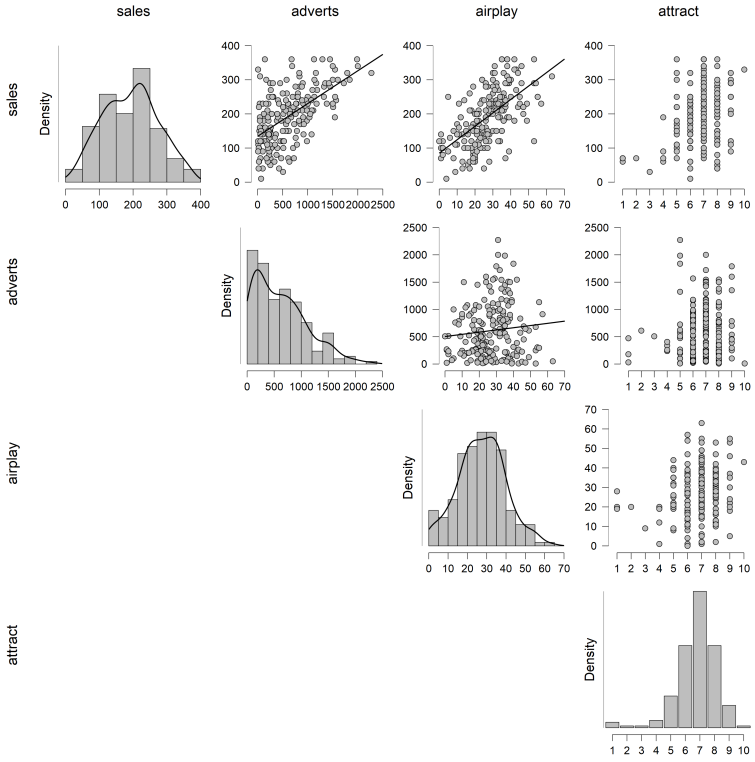

Figure 1

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

							95% CI		Collinearity Statistics	
Model		Unstandardized	Standard Error	Standardized	t	p	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.

Solution.

(a) Model 0: sales = $\beta_0 + \beta_1$ adverts + ε.
Model 1: sales = $\beta_0 + \beta_1$ adverts + β_2 airplay + β_3 attract + ε.

(b)