STA380.18
Homework on Sufficiency, SAS, and Statistics Review
- **Akankshi Mody am92786 (MSBA Class of 2020)**

5. **[7 points]** A new outlet is being planned with an allocated budget of $50,000 per month for advertising and planned average price of $10.00 per pizza. Estimate the monthly sales (number of pizzas) of this outlet. *[Hint: You may want that extra observation in the data set to answer this.]*

The REG Procedure
Model: MODEL1
Dependent Variable: Quantity_Sold Quantity_Sold

Number of Observations Read	16
Number of Observations Used	15
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2206867853	1103433926	59.96	<.0001
Error	12	220828147	18402346		
Corrected Total	14	2427696000			

Root MSE	4289.79552	R-Square	0.9090
Dependent Mean	63740	Adj R-Sq	0.8939
Coeff Var	6.73015		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	817.61031	17704	0.05	0.9639
Average_Price	Average_Price	1	-2617.59327	1189.74351	-2.20	0.0481
Monthly_Advertising_Expenditures	Monthly_Advertising_Expenditures	1	1.69157	0.16608	10.19	<.0001

Quantity_Sold = Intercept + beta0 * Average_Price + beta1 * Monthly_Advertising_Expenditures

= 817.61031 – 2617.59327 * 10 + 1.69157 * 50000

= 59220.17761

= ~ 59,220

6. **[7 points]** By how much, plus or minus, do you expect your estimate in the preceding question will miss actual monthly sales for the outlet? *[Hint: You should be as precise as you can be for this amount.]*

The RMSE can tell us how much we can expect our estimate to deviate. The RMSE for the above estimation is 4,289.79552. Our above estimate can be expected to deviate by ~ 4,290 units on an average.

7. **[7 points]** Can you give an interval in which you are approximately 95% confident that actual sales for the new outlet will lie? *[Hint: You should be as precise as you can be for this amount.]*

The CLI output below shows us that our estimation can range from 48587 to 69853. (This result is achieved after adding the extra observation. The regression is still based on 15 values since SAS ignores the row with the missing value)

The REG Procedure
Model: MODEL1
Dependent Variable: Quantity_Sold Quantity_Sold

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	85300	83889	2344	78783	88995	73238	94539	1411
2	40500	44737	2500	39291	50183	33920	55555	-4237
3	61800	67669	1698	63969	71368	57616	77721	-5869
4	50800	46232	2156	41536	50929	35772	56692	4568
5	60600	54410	1566	50997	57823	44460	64360	6190
6	79400	76855	2308	71826	81883	66241	87468	2545
7	71400	65395	1135	62923	67867	55727	75063	6005
8	70700	69364	1224	66697	72031	59644	79084	1336
9	55600	57616	1310	54762	60470	47843	67388	-2016
10	70900	71901	1705	68186	75617	61843	81960	-1001
11	77200	81799	2014	77410	86187	71473	92124	-4599
12	63200	62175	1668	58540	65810	52146	72204	1025
13	71100	71481	1928	67279	75682	61233	81728	-380.6818
14	55500	54588	2619	48882	60295	43637	65540	911.5834
15	42100	47990	1824	44015	51964	37833	58146	-5890
16	.	59220	2327	54151	64290	48587	69853	.

8. **[7 points]** Suppose the chain adopts a policy that all outlets will spend $50,000 per month on advertising and maintain an average price of $10.00. Estimate the mean monthly sales among all outlets in the chain.

Quantity_Sold = Intercept + beta0 * Average_Price + beta1 * Monthly_Advertising_Expenditures

= 817.61031 – 2617.59327 * 10 + 1.69157 * 50000

= 59220.17761

= ~ 59,220

9. **[7 points]** By how much, plus or minus, do you expect your estimate in the preceding question will miss actual mean monthly sales for the chain? *[Hint: You should be as precise as you can be for this amount.]*

From the CLI output (Q7 answer), we can see that the estimate in the preceding question can deviate from actual mean monthly sales by the standard error = 2327 units.

10. **[7 points]** Can you give an interval in which you are approximately 95% confident that actual mean monthly sales will lie? *[Hint: You should be as precise as you can be for this amount.]*

From the CLI output (Q7 Answer), we can see that the 95% confidence interval for the actual mean sales will be 54151 to 64290.

11. **[7 points]** In a departmental review meeting, the marketing manager stated her opinion that every additional monthly dollar spent on advertising results in an increase in mean sales of about two additional pizzas per month if the average price is not changed. How many standard errors are between her opinion and the corresponding data estimate? Do you think her opinion is consistent with these data, if being within +/- 2 standard errors is considered consistent?

From results shown in Q1, coefficient estimate for monthly advertising expenditure is 1.69. This can be interpreted for $1 extra spend in monthly advertising expenditure given the average price is kept constant, there is a 1.69 units increase in sales. Standard error for the coefficient estimate for monthly expenditure is 0.166. Thus, the range of monthly expenditure estimate with +/- 2 std error is (1.69 +/- 2* 0.166) = 1.358 to 2.022. The number of standard errors between marketing manager's estimate and the data estimate is, (2-1.69)/0.166 = 1.85. Hence her opinion for the estimate of sales of 2 additional pizzas per month is consistent with the data.

12. **[7 points]** In a departmental review meeting, the marketing manager stated her opinion that every additional monthly dollar spent on advertising results in an increase in mean sales of about two additional pizzas per month. How many standard errors are between her opinion and the corresponding data estimate? Do you think her opinion is consistent with these data, if being within +/- 2 standard errors is considered consistent? *[Hint: This is **not** a repeat of the preceding question.]*

Here, we need to run a regression with just monthly advertising expenditure against sales. From the results of this regression, we see that the coefficient estimate for monthly advertising expenditure is 1.75476 with a standard error of 0.18617. Thus, the range of parameter estimate with +/- 2 standard error is (1.75476 +/- 2*0.18617) = (2.13 to 1.38). The number of standard errors between marketing manager's estimate and the data estimate is, (2-1.75)/0.186 = 1.31. Hence her opinion for the estimate of sales of 2 additional pizzas per month is consistent with the data.

The REG Procedure
Model: MODEL1
Dependent Variable: Quantity_Sold Quantity_Sold

Number of Observations Read	16
Number of Observations Used	15
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2117789777	2117789777	88.84	<.0001
Error	13	309906223	23838940		
Corrected Total	14	2427696000			

Root MSE	4882.51372	R-Square	0.8723
Dependent Mean	63740	Adj R-Sq	0.8625
Coeff Var	7.66005		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-32655	10305	-3.17	0.0074
Monthly_Advertising_Expenditures	Monthly_Advertising_Expenditures	1	1.75476	0.18617	9.43	<.0001

13. **[7 points]** Test normality of the residuals in the regression that you ran to answer the preceding question. Use significance level of 0.05.

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.945235	Pr < W	0.4528
Kolmogorov-Smirnov	D	0.150842	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.044276	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.297925	Pr > A-Sq	>0.2500

Shapiro-Wilk: This test is used to check if the population is normally distributed (null hypothesis -> data is normally distributed). Here the p value is 0.4528 (>0.05) and hence the null hypothesis is accepted and the data is normally distributed.

Kolmogorov-Smirnov: This test returns the area that differs from the normal curve. Since the area here is about 15% with a p value of 0.15 which is greater than 0.05. Hence the residuals are normally distributed as per this test.

Cramer-von Mises: The difference for our data is 0.04 depicting the null hypothesis of residuals being normal is accepted.

Anderson-Darling: Area difference from this test is 29.7% with a p value of 0.25, again rejecting the null hypothesis, indicating that the residuals are normal.