MATH 4044 – Statistics for Data Science ——Cou Assignment 3

(a)Use SAS to perform a one-way ANOVA test to determine whether there is a statistically significant difference in the mean number of boxes sold during the promotion period, by type of promotion:

- **Check the necessary conditions;**
- **Examine and comment on residuals;**
- **If appropriate, perform post-hoc tests;**
- **Report and briefly discuss your results.**

First of all, through the descriptive statistics in the table below, we can find that when tea=3, whether it is the average, the minimum or the maximum, it is lower than the other two cases, this difference may be proved to have statistically significant.

Descriptive statistics

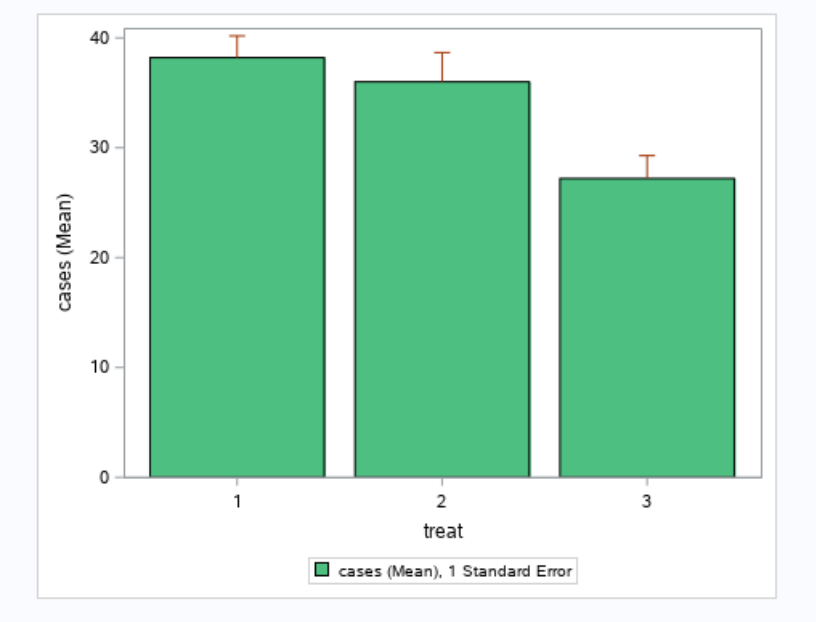
The MEANS Procedure

Analysis Variable : cases						
treat	N Obs	N	Mean	Std Dev	Minimum	Maximum
1	5	5	38.200	4.438	33.000	45.000
2	5	5	36.000	5.958	27.000	43.000
3	5	5	27.200	4.658	21.000	32.000

By observing the figure below, it is not difficult to find that there are differences in the average value of treat under different circumstances. When treat=3, the average value is significantly lower than the other two.

Lacking some results: covariate b/w cases and last
ANOVA between last and treat to check independence.
ANCOVA with interaction term to check equal slope.
residuals
Many incorrect interpretations.
Need to revise the ANCOVA procedure.

10



We then performed a normality analysis for the three cases of treat, and the results are as follows:

Normality check

The UNIVARIATE Procedure
Variable: cases
treat = 1

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.960478	Pr < W	0.8113
Kolmogorov-Smirnov	D	0.228481	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.035413	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.226318	Pr > A-Sq	>0.2500

Normality check

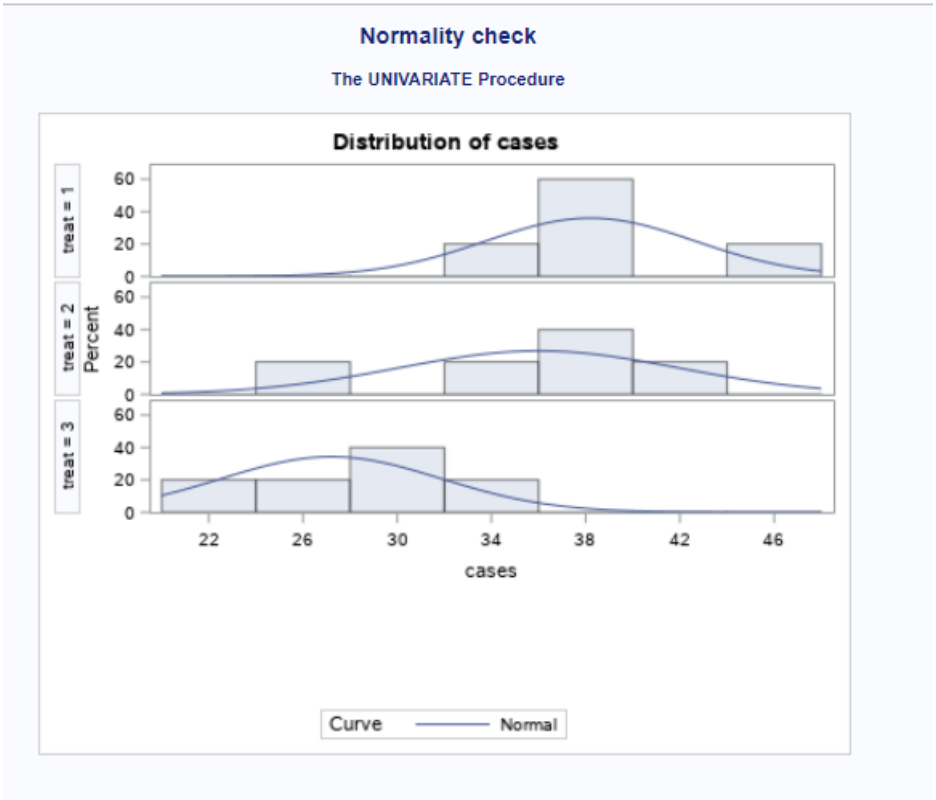
The UNIVARIATE Procedure
Variable: cases
treat = 2

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.947551	Pr < W	0.7197
Kolmogorov-Smirnov	D	0.231441	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.044939	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.264286	Pr > A-Sq	>0.2500

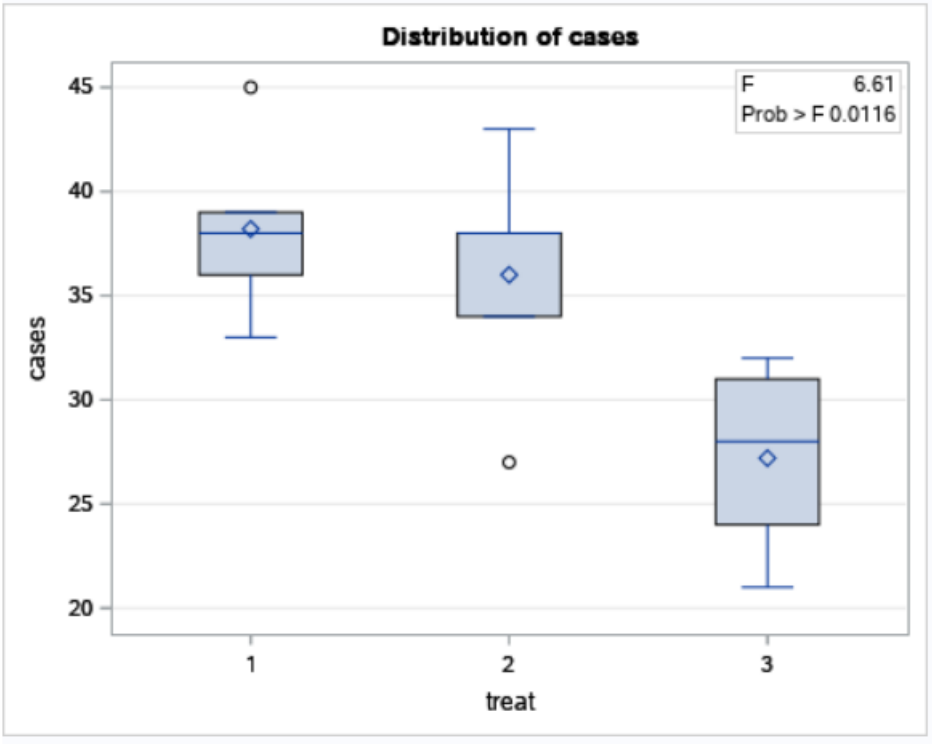
Normality check

The UNIVARIATE Procedure
Variable: cases
treat = 3

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.933294	Pr < W	0.6190
Kolmogorov-Smirnov	D	0.192677	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.035527	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.235216	Pr > A-Sq	>0.2500



The above histogram shows the distribution of cases divided by treat. From the histogram, no matter what the value of treat is, the distribution of cases does not show a normal distribution. However, from the point of view of the p-value, the p-value of the three is greater than 0.05, so we cannot reject the hypothesis of normal distribution.



The boxplot can support the same conclusion. When treat is 1 and 2, there are outliers in cases, and the asymmetry in the distribution is obvious, but when treat=3, there are no outliers, and the distribution is approximately normal distribution, or there is a little skewed distribution.

Let's try the one-way anova test. Here are the results of the test:

Normality check

The GLM Procedure

Class Level Information		
Class	Levels	Values
treat	3	1 2 3

Number of Observations Read	15
Number of Observations Used	15

Normality check

The GLM Procedure

Dependent Variable: cases

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	338.8000000	169.4000000	6.61	0.0116
Error	12	307.6000000	25.6333333		
Corrected Total	14	646.4000000			

R-Square	Coeff Var	Root MSE	cases Mean
0.524134	14.97910	5.062937	33.80000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
treat	2	338.8000000	169.4000000	6.61	0.0116

Source	DF	Type III SS	Mean Square	F Value	Pr > F
treat	2	338.8000000	169.4000000	6.61	0.0116

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	27.20000000	B	2.26421436	12.01	<.0001
treat 1	11.00000000	B	3.20208266	3.44	0.0049
treat 2	8.80000000	B	3.20208266	2.75	0.0177
treat 3	0.00000000	B	.	.	.

Before interpreting the table, we first need to check the results of the homogeneity of variance test in the table. There are significant differences in variance across the three treats, F(2,12) = 6.61, P-value = 0.0116 ≈ 0.01.

wrong conclusion

Due to the violation of the assumption of homogeneity of variances, we report the Welch-adjusted F-ratio shown in the table instead of the one given in the main ANOVA table.

what is your conclusion from Welch then?

The table also indicates the situation relative to the average value when treat=3. The parameter treat = 1 was statistically significant (p-value < 0.01), the data at treat = 2 required further analysis because the p-value was 0.0177. The intercept of 27.2 represents the sample mean for treat=1, and the other parameters show the difference between the mean of the other treat values and the values for treat=1.

why need further?

Are the difference statisticallly significant?

Normality check

The GLM Procedure

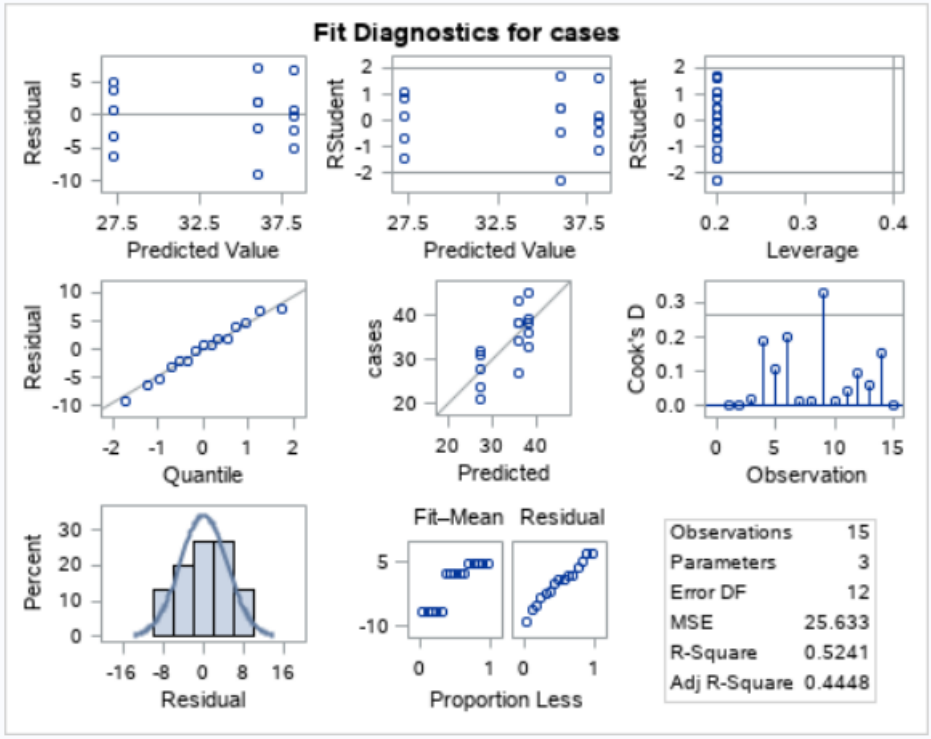
Levene's Test for Homogeneity of cases Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
treat	2	473.7	236.8	0.38	0.6906
Error	12	7443.9	620.3		

no difference in variance.

Welch's ANOVA for cases			
Source	DF	F Value	Pr > F
treat	2.0000	7.20	0.0166
Error	7.8923		

p<0.05, hence significant difference.

Since Welch's F(2,7.8923) = 7.20 with P-value > 0.01, there is no significant difference among mean by treat.



Studentised residuals plot shows a few points outside the -2 and 2 bounds, but most are close to the bounds and the number is not too large (less than 5%).

what do you mean by this? p-value from QQ?

From the results of the Q-Q plot, the observed value of the p value is the same as the expected value, indicating that the analytical model is reasonable. However, all the observed values of P value did not significantly exceed the expected value, indicating that the analysis results did not find a significantly associated locus (with the trait). Possible reasons include: the trait is controlled by a small polygene, the

what?

effect is too weak; the population size is not enough, etc. From the histogram, the residuals are slightly skewed, but not very noticeable. no skew here.

The R-squared is 0.5241, which is a very high value.

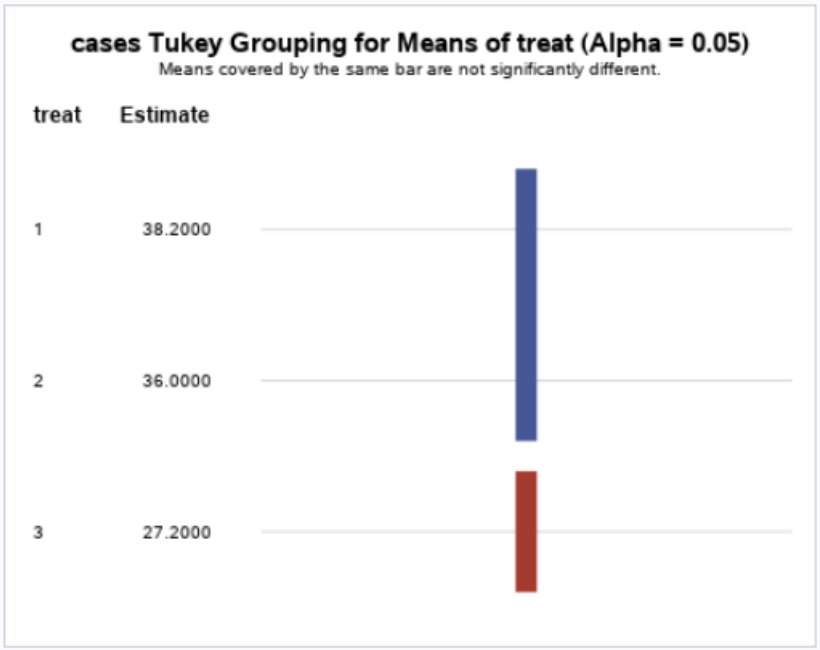
Normality check

The GLM Procedure

Tukey's Studentized Range (HSD) Test for cases

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	12
Error Mean Square	25.63333
Critical Value of Studentized Range	3.77289
Minimum Significant Difference	8.5426



Normality check

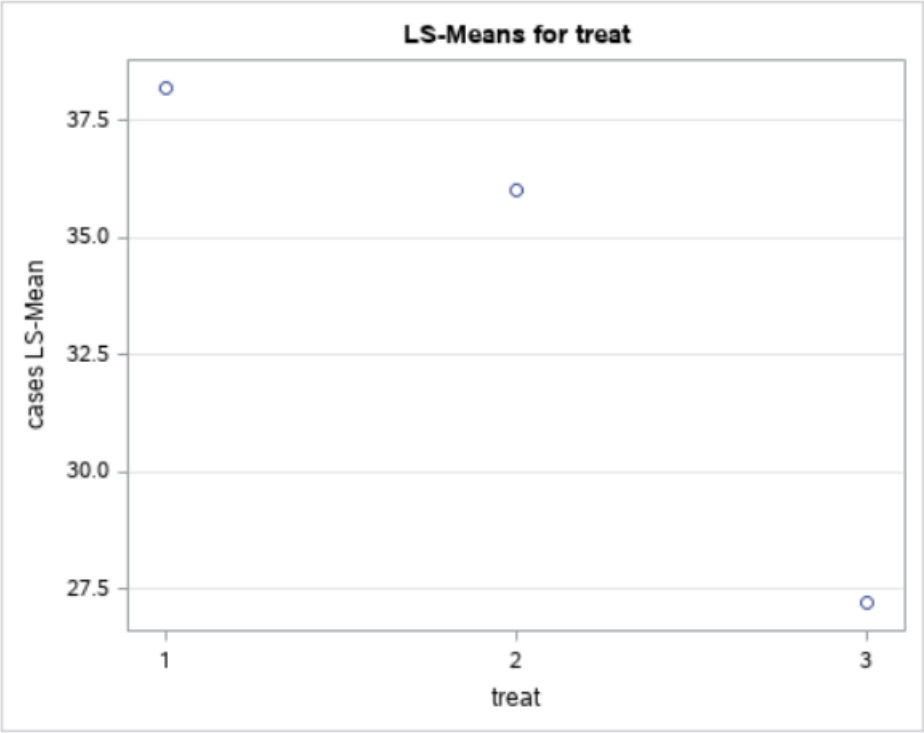
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

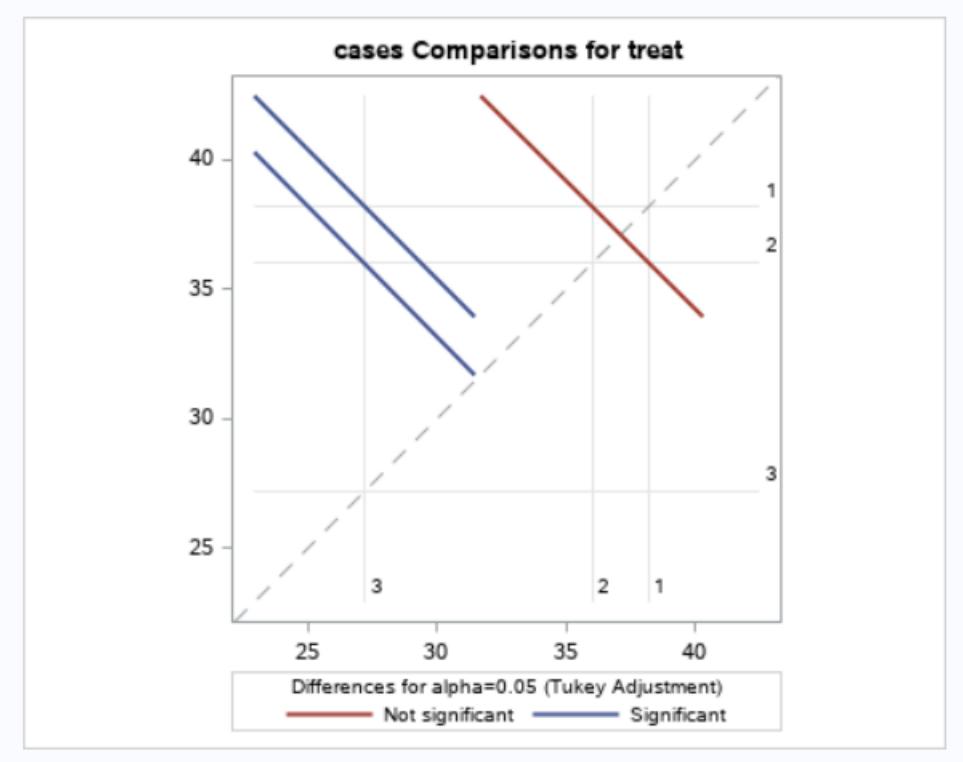
treat	cases LSMEAN	LSMEAN Number
1	38.2000000	1
2	36.0000000	2
3	27.2000000	3

Least Squares Means for effect treat
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: cases

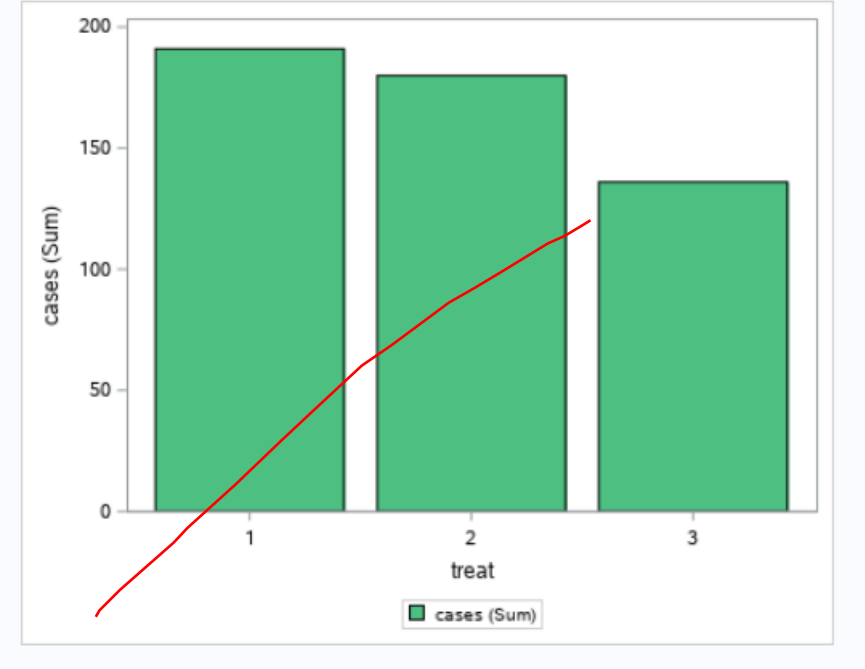
i/j	1	2	3
1		0.7753	0.0127
2	0.7753		0.0434
3	0.0127	0.0434	





From the above figure, we can think that there is no significant difference when treat=1 and treat-2, and there are significant differences in other cases.

(b) Use SAS to perform a one-way ANCOVA with the number of the cases sold in the preceding period (variable last) as a covariate:
- Confirm that there is a linear relationship between the response variable and the covariate (a scatterplot and a correlation coefficient plus a comment will suffice);
- Check the two additional ANCOVA assumptions (report and comment only on the parts of the output most directly relevant to condition checking for this exercise):
- Independence of the covariate and the treatment effect (perform a one-way ANOVA test; there should be no statistically significant difference);
- Equality of slopes (add and check significance of the interaction term);
- Decide what your final ANCOVA model should be (with or without the interaction term) and perform post-hoc analysis for this model;
- Examine and comment on residuals;
- Report and briefly discuss your results.
Note: You should obtain and examine Type III Sum of Squares (ss3). Also obtain estimates of 'least squares means' (lsmeans) which are means adjusted for the covariate.



By observing the histogram, it is not difficult to find that when the value of treat is 3, the total sales of cases are significantly lower than the two cases.

The GLM Procedure

Class Level Information		
Class	Levels	Values
treat	3	1 2 3

Number of Observations Read	15
Number of Observations Used	15

The GLM Procedure

Dependent Variable: cases

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	607.8286915	202.6095638	57.78	<.0001
Error	11	38.5713085	3.5064826		
Corrected Total	14	646.4000000			

R-Square	Coeff Var	Root MSE	cases Mean
0.940329	5.540120	1.872560	33.80000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
treat	2	417.1509137	208.5754568	59.48	<.0001
last	1	269.0286915	269.0286915	76.72	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	4.37659064	B	2.73692149	1.60	0.1381
treat 1	12.97683073	B	1.20562330	10.76	<.0001
treat 2	7.90144058	B	1.18874585	6.65	<.0001
treat 3	0.00000000	B	.	.	.
last	0.89855942		0.10258488	8.76	<.0001

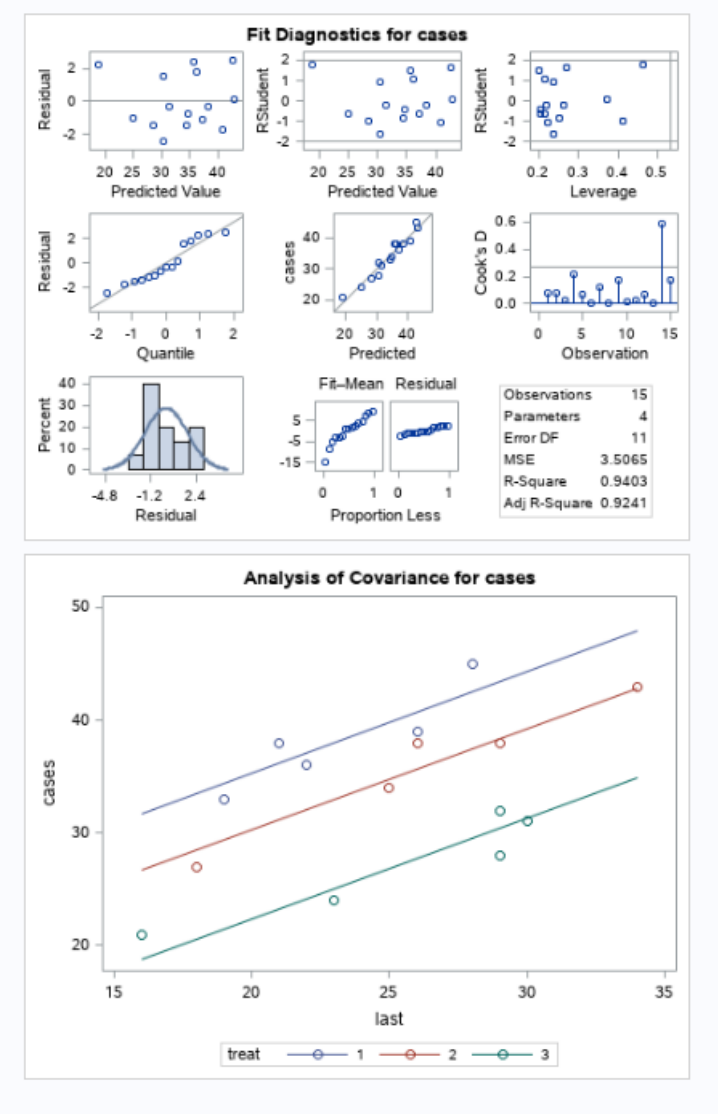
Before interpreting the table, we first need to check the results of the homogeneity of variance test in the table. There are significant differences in variance across the three treats, F(3,11) = 57.78, P-value < 0.0001.

NOt relevant for ANCOVA.

Due to the violation of the assumption of homogeneity of variances, we report the Welch-adjusted F-ratio shown in the table instead of the one given in the main ANOVA table.

The table also indicates the situation relative to the total value when treat=3. The parameter treat = 1 was statistically significant (p-value < 0.01), the data at treat = 2 required further analysis because the p-value was 0.0177. The intercept of 4.37659 represents the sample total for treat=3, and the other parameters show the difference between the total of the other treat values and the values for treat=3.

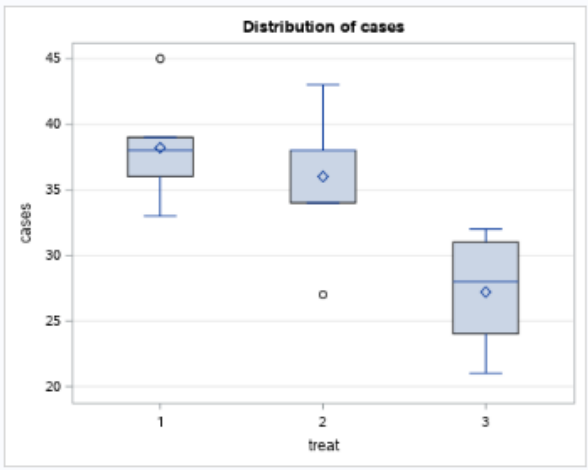
wrong interpretation of the model equation.



Studentised residuals plot shows a few points outside the -2 and 2 bounds, but most are close to the bounds and the number is not too large (less than 5%).

Judging from the Q-Q plot and the histogram of the residuals, we can completely deny the normal distribution assumption, because this is an obvious positive skewed distribution.

The R-squared is 0.9403, which is a very high value.

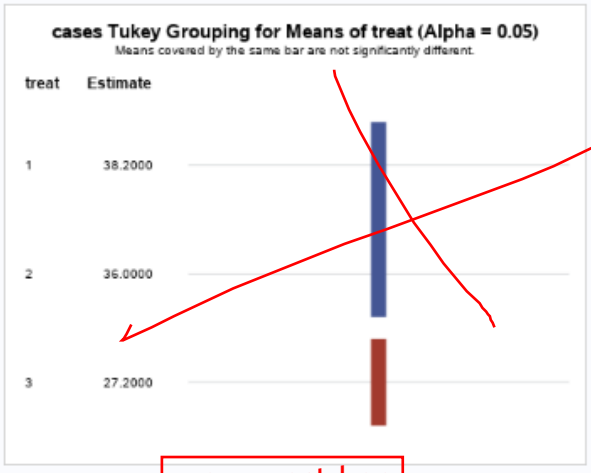


The GLM Procedure

Tukey's Studentized Range (HSD) Test for cases

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	11
Error Mean Square	3.506483
Critical Value of Studentized Range	3.81958
Minimum Significant Difference	3.1986



wrong post-hoc

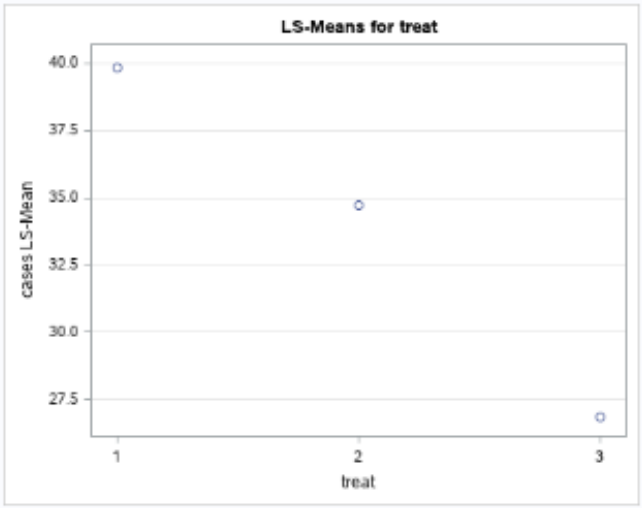
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

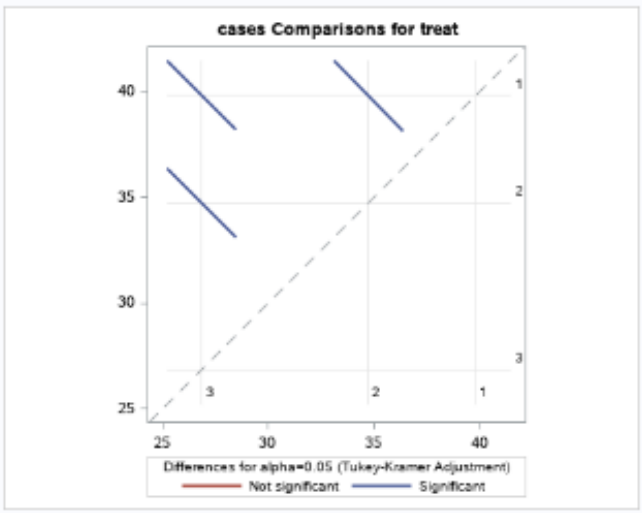
treat	cases LSMEAN	LSMEAN Number
1	39.8174070	1
2	34.7420168	2
3	26.8405762	3

Least Squares Means for effect treat
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: cases

i/j	1	2	3
1		0.0044	<.0001
2	0.0044		<.0001
3	<.0001	<.0001	





From the graph above, we can think that there is no significant difference in all three cases.

(c)Compare results from parts (a) and (b). Did including the covariate reduce the error variance and thus produce better estimates of mean sales levels by the type of promotion? Which model is a better fit to the data? Which type of promotion appears to be the most effective? Explain briefly.

From an analytical point of view, we have sufficient reasons to believe that when treat=1, the sales situation is the best.