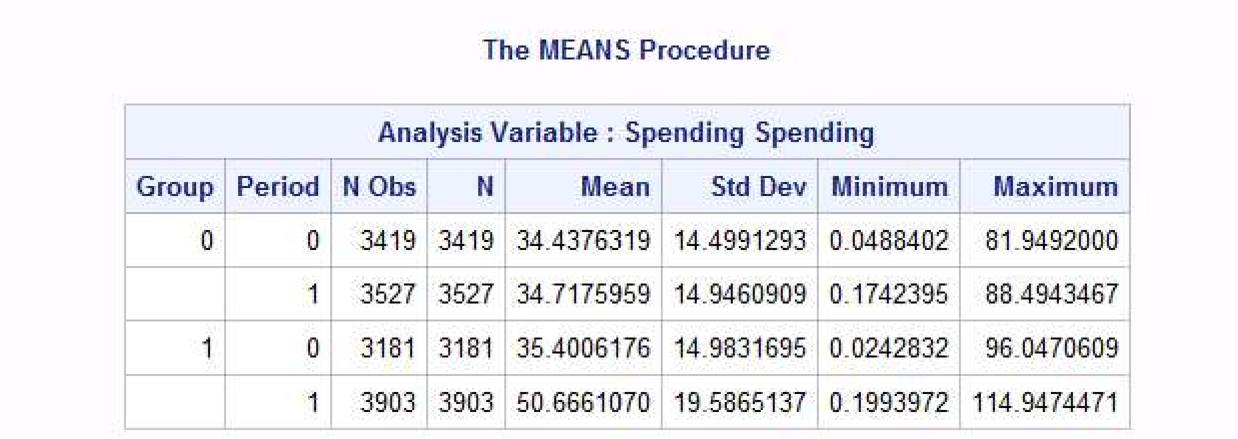
Questions

1. It is instructive to check if there are prima facie differences between the consumer groups with respect to their social media participation through some summary statistics. Let us analyze consumer spending and see how it changes across the consumer group vis-à-vis their social media participation. Consider the following tabulation of consumer spending (this is equivalent to using cross tabs that you have encountered in Marketing Research).

	Mean or Average Weekly Spending		
	Participating or Treatment Customers	Non-Participating or Control Customers	Difference
Before Social Media Launch	$\bar{X}_1$ 10 ($StdDev_1$)	$\bar{X}_3$ 00 ($StdDev_3$)	D_3
After Social Media Launch	$\bar{X}_2$ 11 ($StdDev_2$)	$\bar{X}_4$ 01 ($StdDev_4$)	D_4
Difference	D_1	D_2	

- Calculate all the values in the table, i.e., $\bar{X}_1$, $StdDev_1$, D_1 and the rest. Based on the values, what do you find (summarize your findings in about 150-200 words).



The MEANS Procedure

Analysis Variable : Spending Spending							
Group	Period	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	0	3419	3419	34.4376319	14.4991293	0.0488402	81.9492000
	1	3527	3527	34.7175959	14.9460909	0.1742395	88.4943467
1	0	3181	3181	35.4006176	14.9831695	0.0242832	96.0470609
	1	3903	3903	50.6661070	19.5865137	0.1993972	114.9474471

As we can see from the above table obtained from the MEANS procedure the mean and standard deviation is given below:

- Participating customers before social media launch (10)

  $\bar{X}_1$ (Mean) = 35.4006176

  $StdDev_1$ = 14.9831695

- Participating customers after social media launch (11)

  $\bar{X}_2$ (Mean) = 50.6661070

$StdDev_2$ = 19.5865137

- Non-participating customers before social media launch (00)

  $\overline{X}_3$ (Mean) = 34.4376319

  $StdDev_3$ = 14.4991293

- Non-participating customers after social media launch (01)

  $\overline{X}_4$ (Mean) = 34.7175959

  $StdDev_4$ = 14.9460909

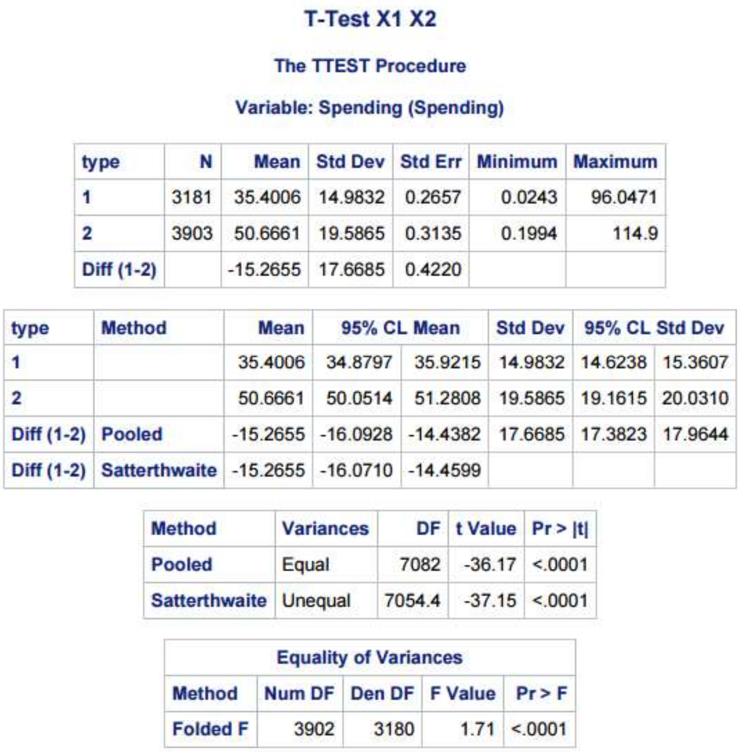
- As expected the mean is highest for the section where group 1 and period 1, i.e., participating customers after social media launch.

The difference values can be summarized as below:

- The difference between the spending of participating customers before and after the social media launch is D_1 which has a value of 15.2654894.
- The difference between the spending of non-participating customers before and after the social media launch is D_2 which has a value of 0.279964.
- The difference between the spending of participating and non-participating customers before the social media launch is D_3 which has a value of 0.9629857.
- The difference between the spending of participating and non-participating customers after the social media launch is D_4 which has a value of 15.9485111.

  Therefore, we can conclude that the social media launch has a positive influence on the participating customers spending behavior whereas, the effect it has on the non-participating customers is not significant.

- In order to justify your findings, perform the appropriate paired t-tests. That is you will test if the pairs $\overline{X}_1$ and $\overline{X}_2$; $\overline{X}_3$ and $\overline{X}_4$; $\overline{X}_1$ and $\overline{X}_3$; $\overline{X}_3$ and $\overline{X}_4$ are statistically different from each other (this of course is equivalent to testing if D_1, D_2, D_3, D_4 each is different from zero). Summarize your findings in about 100 words.



T-Test X1 X2

The TTEST Procedure

Variable: Spending (Spending)

type	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	3181	35.4006	14.9832	0.2657	0.0243	96.0471
2	3903	50.6661	19.5865	0.3135	0.1994	114.9
Diff (1-2)		-15.2655	17.6685	0.4220		

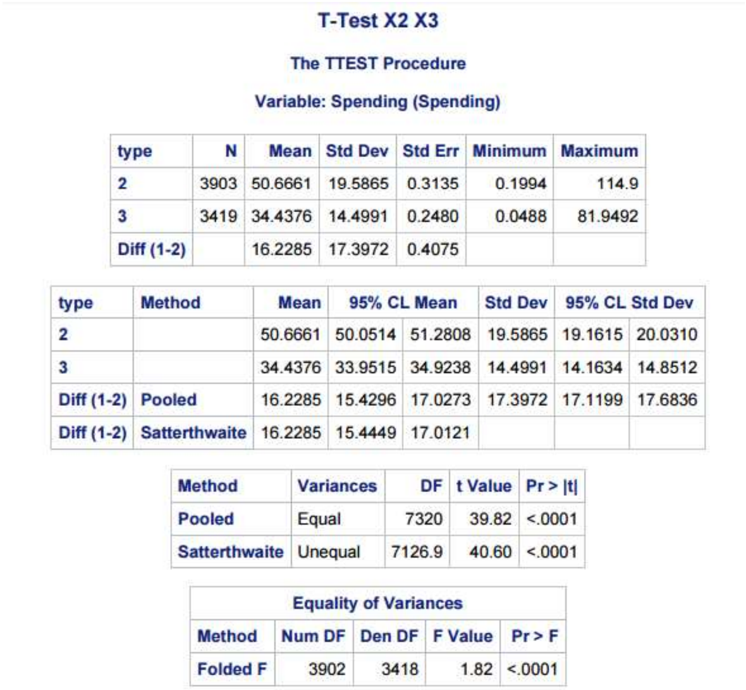
type	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		35.4006	34.8797	35.9215	14.9832	14.6238	15.3607
2		50.6661	50.0514	51.2808	19.5865	19.1615	20.0310
Diff (1-2)	Pooled	-15.2655	-16.0928	-14.4382	17.6685	17.3823	17.9644
Diff (1-2)	Satterthwaite	-15.2655	-16.0710	-14.4599			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	7082	-36.17	<.0001
Satterthwaite	Unequal	7054.4	-37.15	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	3902	3180	1.71	<.0001

The t-test between X1 and X2:

- We observe from the F statistic that the p-value is less that α and therefore, we take into account unequal variances and use the Satterthwaite method with a difference of 15.2655.



T-Test X2 X3

The TTEST Procedure

Variable: Spending (Spending)

type	N	Mean	Std Dev	Std Err	Minimum	Maximum
2	3903	50.6661	19.5865	0.3135	0.1994	114.9
3	3419	34.4376	14.4991	0.2480	0.0488	81.9492
Diff (1-2)		16.2285	17.3972	0.4075		

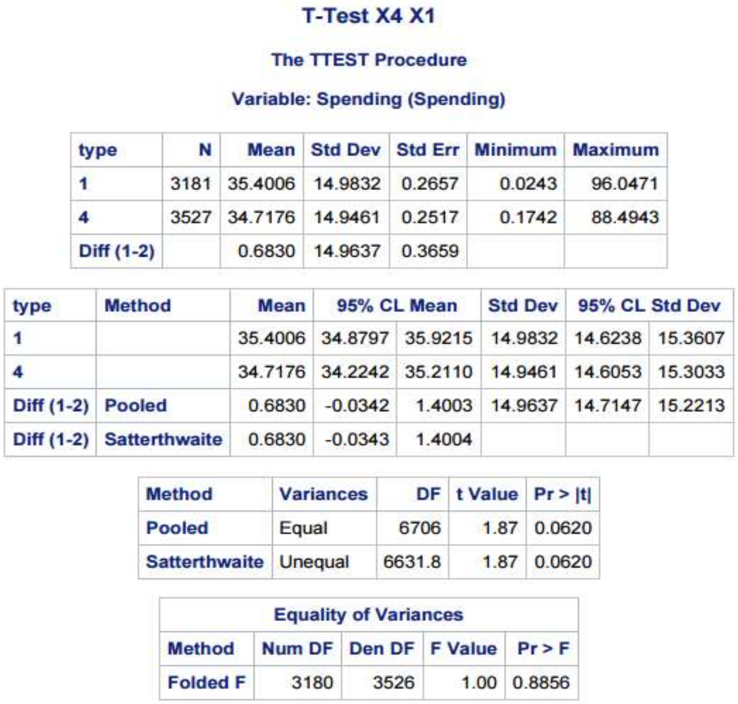
type	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
2		50.6661	50.0514	51.2808	19.5865	19.1615	20.0310
3		34.4376	33.9515	34.9238	14.4991	14.1634	14.8512
Diff (1-2)	Pooled	16.2285	15.4296	17.0273	17.3972	17.1199	17.6836
Diff (1-2)	Satterthwaite	16.2285	15.4449	17.0121			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	7320	39.82	<.0001
Satterthwaite	Unequal	7126.9	40.60	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	3902	3418	1.82	<.0001

The t-test between X3 and X2:

- We observe from the F statistic that the p-value is less that α and therefore, we take into account unequal variances and use the Satterthwaite method with a difference of 16.2285.



T-Test X4 X1

The TTEST Procedure

Variable: Spending (Spending)

type	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	3181	35.4006	14.9832	0.2657	0.0243	96.0471
4	3527	34.7176	14.9461	0.2517	0.1742	88.4943
Diff (1-2)		0.6830	14.9637	0.3659		

type	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		35.4006	34.8797	35.9215	14.9832	14.6238	15.3607
4		34.7176	34.2242	35.2110	14.9461	14.6053	15.3033
Diff (1-2)	Pooled	0.6830	-0.0342	1.4003	14.9637	14.7147	15.2213
Diff (1-2)	Satterthwaite	0.6830	-0.0343	1.4004			

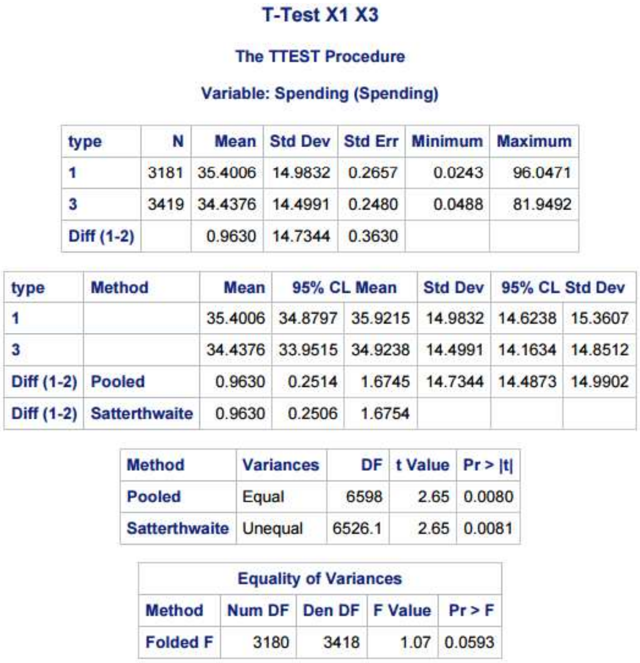
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	6706	1.87	0.0620
Satterthwaite	Unequal	6631.8	1.87	0.0620

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	3180	3526	1.00	0.8856

The t-test between X4 and X1:

- We observe from the F statistic that the p-value is greater that α and therefore, we take into account equal variances and use the Pooled method with a difference of 0.6830.



T-Test X1 X3

The TTEST Procedure

Variable: Spending (Spending)

type	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	3181	35.4006	14.9832	0.2657	0.0243	96.0471
3	3419	34.4376	14.4991	0.2480	0.0488	81.9492
Diff (1-2)		0.9630	14.7344	0.3630		

type	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		35.4006	34.8797	35.9215	14.9832	14.6238	15.3607
3		34.4376	33.9515	34.9238	14.4991	14.1634	14.8512
Diff (1-2)	Pooled	0.9630	0.2514	1.6745	14.7344	14.4873	14.9902
Diff (1-2)	Satterthwaite	0.9630	0.2506	1.6754			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	6598	2.65	0.0080
Satterthwaite	Unequal	6526.1	2.65	0.0081

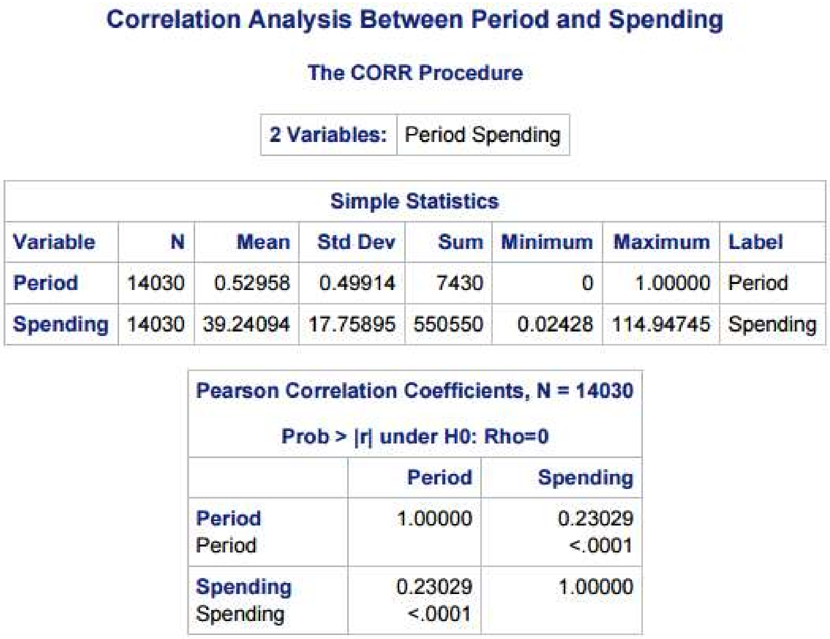
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	3180	3418	1.07	0.0593

The t-test between X1 and X3:

- We observe from the F statistic that the p-value is greater that α and therefore, we take into account equal variances and use the Pooled method with a difference of 0.9630.

Therefore, the paired t-test confirms our conclusion made earlier, i.e., social media launch has positively impacted the participating customers spending behavior the most.

- Now assume that someone tells you that you can get the same information by testing the correlations of the paired variables above. Run the correlation analysis for the different pairs of variables. Is this a good test/method for analyzing the above, why or why not?



Correlation Analysis Between Period and Spending

The CORR Procedure

2 Variables:	Period Spending

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Period	14030	0.52958	0.49914	7430	0	1.00000	Period
Spending	14030	39.24094	17.75895	550550	0.02428	114.94745	Spending

Pearson Correlation Coefficients, N = 14030 Prob > |r| under H0: Rho=0		
	Period	Spending
Period Period	1.00000	0.23029 <.0001
Spending Spending	0.23029 <.0001	1.00000

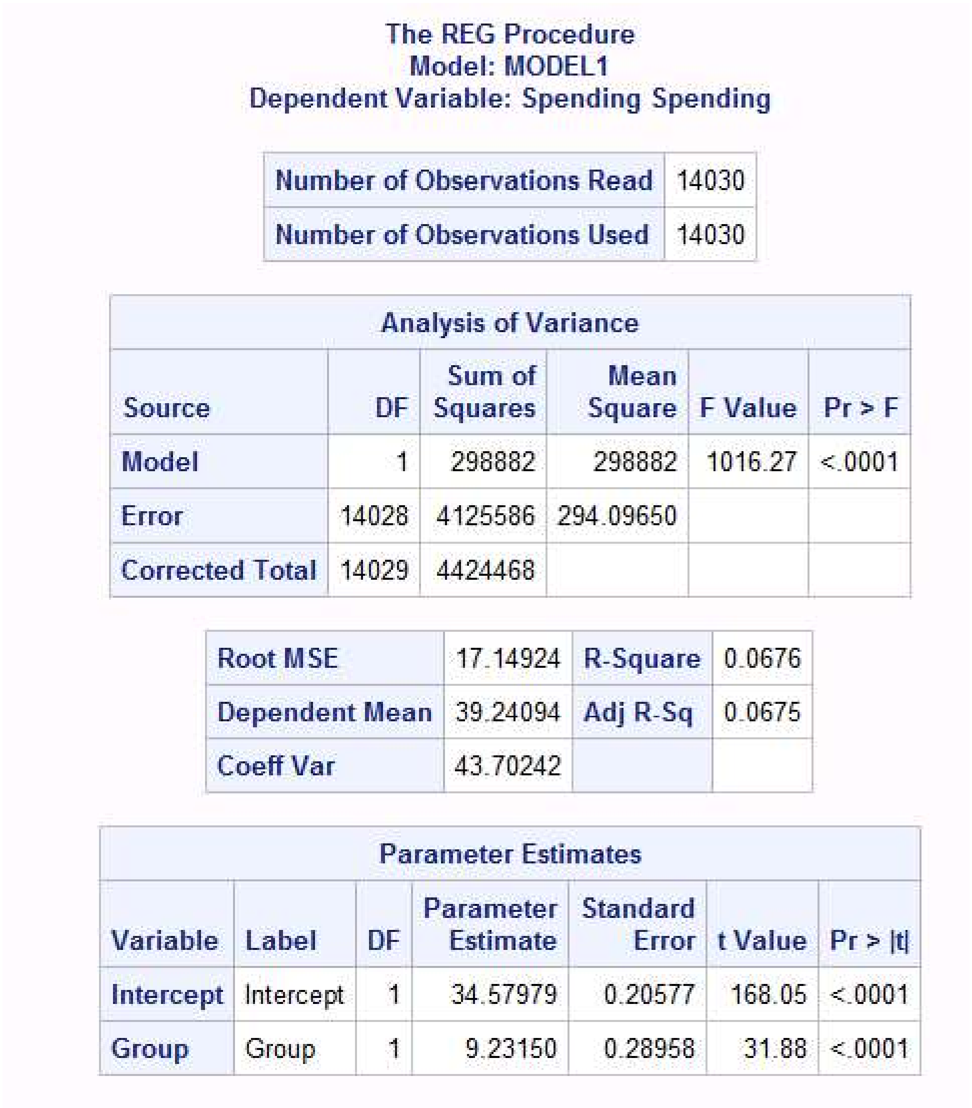
We ran the correlation analysis for period and spending and according to us this method is not an efficient way to find the relationship between the dependent and the independent variables due to the fact that the independent variable is binary and correlation cannot be effectively applied on this.

2. Now assume that you are asked to run a OLS (ordinary Least Squares) regression model to analyze the impact of social media participation as follows:

$$Spending_i = \beta_0 + \beta_1 Group_i + \varepsilon_i$$

where spending is the a the amount that the consumers spend on the firm's products that week and group takes the value of 1if eventually the consumer has become part of the firms' social media page and 0 otherwise (as indicated earlier). Run the above regression and summarize your findings.

- Is the above a correct regression/model to analyze the impact of consumers social media participation on their spending? Discuss with appropriate reasoning in about 150-200 words.



The REG Procedure
Model: MODEL1
Dependent Variable: Spending Spending

Number of Observations Read	14030
Number of Observations Used	14030

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	298882	298882	1016.27	<.0001
Error	14028	4125586	294.09650		
Corrected Total	14029	4424468			

Root MSE	17.14924	R-Square	0.0676
Dependent Mean	39.24094	Adj R-Sq	0.0675
Coeff Var	43.70242		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	34.57979	0.20577	168.05	<.0001
Group	Group	1	9.23150	0.28958	31.88	<.0001

- As we can see from the above results, the p-values for both group and intercept is less than α, we consider both these values in the regression equation.

- Therefore, the regression equation obtained for the above scenario can be modelled as below:

  Spending = 34.57979 + 9.23150 Group

- In terms of interpretation, a participating group leads to a spending of $43.81129 and a non-participating group leads to a spending of $34.57979

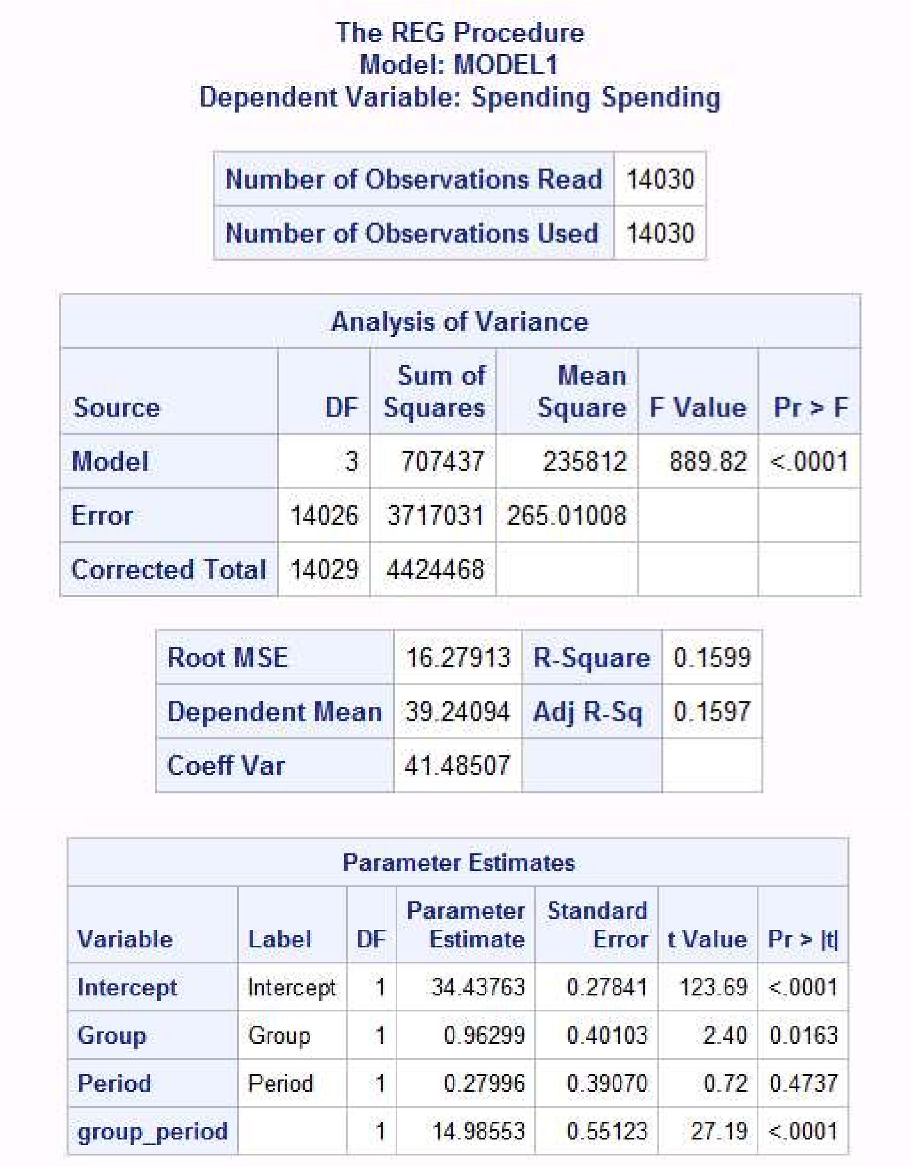
- According to us, this is not an effective model as it does not take into consideration the effect of the independent variable – ‘Period’, which is the variable that differentiates the period prior to the social media launch and the period post the social media launch.

- Therefore, we conclude that in order to get at better estimates we need to consider the period's effect too.

3. A more appropriate regression model is to utilize the information you have about consumer spending before the social media page launch by the company (consumers participating in the social media). You can specify a regression model as follows:

$$Spending_i = \beta_0 + \beta_1 Group_i + \beta_2 Period_i + \beta_3 Group_i \times Period_i + \varepsilon_i$$

where period takes the value of 0 before the firm has launched the social media page (and therefore consumers cannot participate in the firm's social media) and 1 afterwards. Such a model is referred to as the DID or the DD model in the literature.



The REG Procedure
Model: MODEL1
Dependent Variable: Spending Spending

Number of Observations Read	14030
Number of Observations Used	14030

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	707437	235812	889.82	<.0001
Error	14026	3717031	265.01008		
Corrected Total	14029	4424468			

Root MSE	16.27913	R-Square	0.1599
Dependent Mean	39.24094	Adj R-Sq	0.1597
Coeff Var	41.48507		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	34.43763	0.27841	123.69	<.0001
Group	Group	1	0.96299	0.40103	2.40	0.0163
Period	Period	1	0.27996	0.39070	0.72	0.4737
group_period		1	14.98553	0.55123	27.19	<.0001

- Run the above model and report your findings. Interpret your model parameters and discuss them especially focusing on β_3. How do you interpret β_3.
  - Both Group and Period have p-value less than the α- value of 5%.
  - So, the final equal would be;
    - Spending = 34.43763 + 14.98553 Group_Period
  - β_3 - This coefficient implies unit change in group_period leads to a change in spending by $14.98553.

- Compare your model estimates obtained from #3 with your answers from #1. What do you find? Discuss in detail.

  As we can see from the above equation, a participating group's spending after the social media launch equals a value of 49.4236 which is very close to the mean value which is, 50.6661070, obtained for Group 1 and Period 1 spending from the MEANS procedure. This affirms the fact that the model estimate obtained from the above equation is a quite accurate.

- How much does the spending increase/decrease for consumer who participate in social media?

  The spending value for participating consumers before and after social media launch, after plugging in the corresponding values in the above equation is given below:

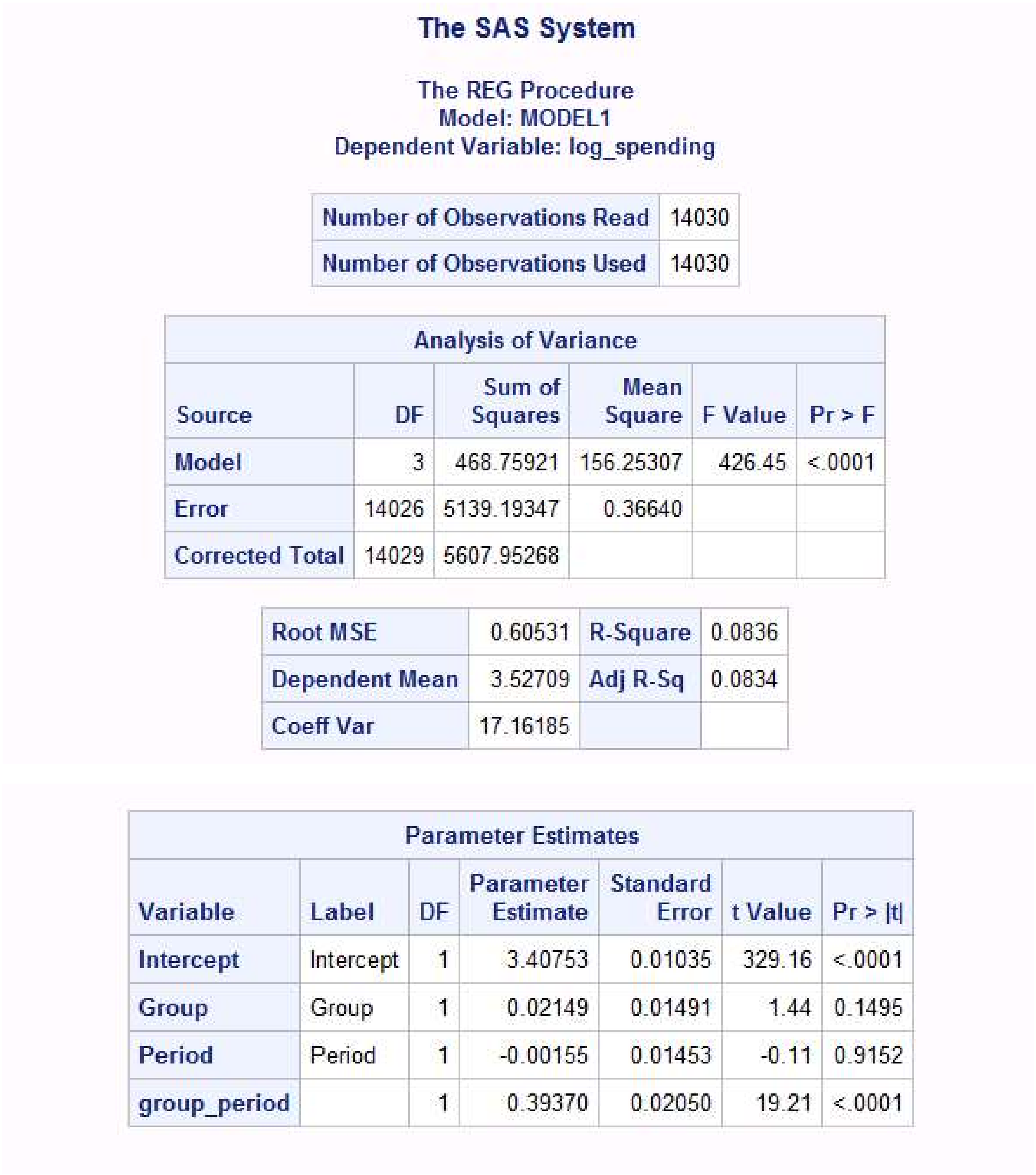
  Participating consumers before the social media launch: 34.43763
  Participating consumers after the social media launch: 49.42316

  As seen from the values obtained above the amount that the participating consumers spend increases by 43.5% which is a very convincing indicator of the positive effect (in terms of monetary value) the launch of the social media webpage has on the consumer purchase behavior.

4. Another way to model the above is to take the logarithm (natural log) of the dependent variable (spending). This model would be:

$$\log(Spending_i) = \beta_0 + \beta_1 Group_i + \beta_2 Period_i + \beta_3 Group_i \times Period_i + \varepsilon_i$$

   What would be the advantage of using such as specification of the dependent variable? How will you interpret the parameters now? Compare your answers of model #4 with that of #3. What is the elasticity of customer social media participation on spending? What does the elasticity imply? How do you interpret the value?



The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: log_spending

Number of Observations Read	14030
Number of Observations Used	14030

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	468.75921	156.25307	426.45	<.0001
Error	14026	5139.19347	0.36640		
Corrected Total	14029	5607.95268			

Root MSE	0.60531	R-Square	0.0836
Dependent Mean	3.52709	Adj R-Sq	0.0834
Coeff Var	17.16185		

Parameter Estimates

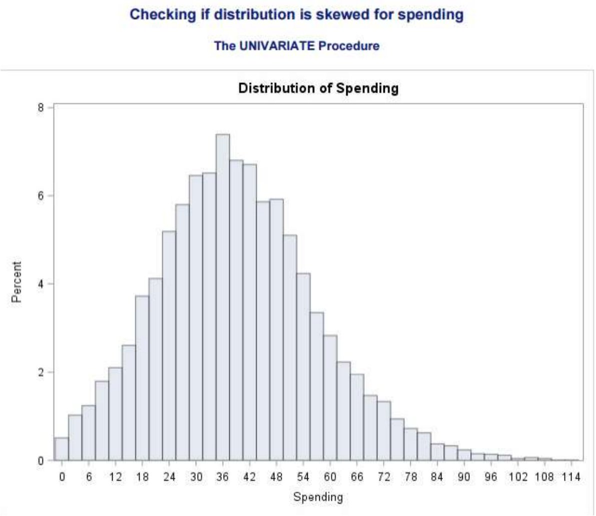
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	3.40753	0.01035	329.16	<.0001
Group	Group	1	0.02149	0.01491	1.44	0.1495
Period	Period	1	-0.00155	0.01453	-0.11	0.9152
group_period		1	0.39370	0.02050	19.21	<.0001

- Both Group and Period have p-value less than the α- value of 5%.
- So, the final equal would be;
  - Log(Spending) = 3.40753 + 0.39370 Group_Period

Taking log transformations is the most common way to handle situations where a non-linear relationship exists between independent and dependent, in other words it helps to model the relationship between the independent and the dependent variable better. The advantages of taking the log of the dependent variable is given below:

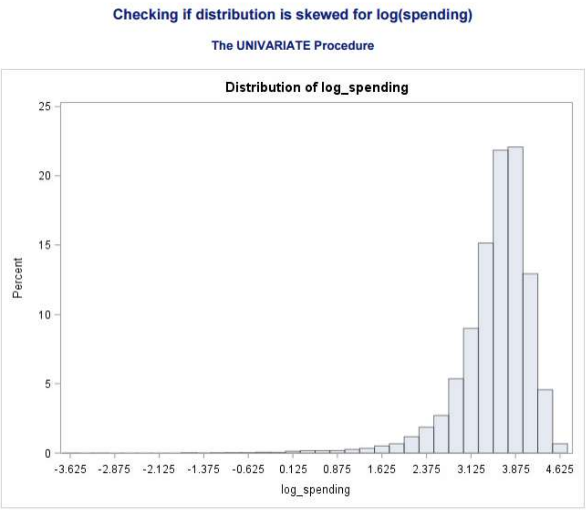
1. The effective relationship is non-linear while preserving the linear relationship.
2. Log transformations is a convenient way of changing skewed distributions into normal.

We ran the ‘proc univariate’ code to find the distribution of the data for both spending and log(spending) and compared the two to find if taking the log of the dependent variable helps remove the skewness. We observed that the ‘Spending’ data was distributed normally and taking the log of spending skewed the data further.



Checking if distribution is skewed for spending

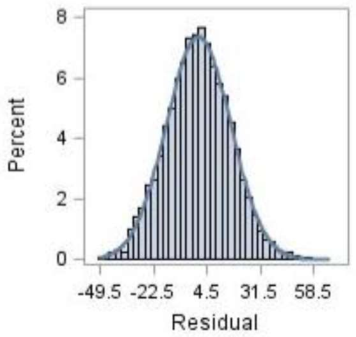
The UNIVARIATE Procedure



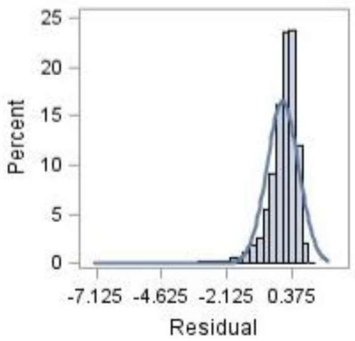
Checking if distribution is skewed for log(spending)

The UNIVARIATE Procedure

Residual Graphs for the two models also confirm that the first model without the log of the dependent variable is better as the residual is distributed normally.



Linear model



Log-Linear model

Therefore, we can conclude that the only advantage of taking the logarithm of the dependent variable is the ease of explanation of the dependent variable,spending, in terms of the independent variables.

Log(*Spending*) = 3.40753 + 0.39370 *Group_Period*

We can interpret the above the equation as – A unit change in group_period would lead to 39.37% increase in spending.

- model #3: For a participating member after social media launch we see that the spending would be 49.42316.
- model #4: For a participating member after social media launch we see that the spending would be 39.37%

Elasticity is defined as relative percentage change of one variable with respect to another and derived as below;

$$\text{Log}(\mathit{Spending}) = 3.40753 + 0.39370\ \mathit{Group\_Period}$$

$$\mathit{Spending} = e^{3.40753+0.39370\ Group\_Period}$$

$$\frac{dSpending}{dGroup\_Period} = 0.39370\ e^{3.40753+0.39370\ Group\_Period}$$

$$\frac{dSpending}{dGroup\_Period} = 0.39370 Spending$$

$$\frac{dSpending}{Spending} = 0.39370 dGroup\_Period$$

$$\frac{dSpending}{Spending} * 100 = 39.370 dGroup\_Period$$

and the elasticity is given by:

$$e = \frac{dy}{dx} . \frac{x}{y} = \beta x$$

$$e = \frac{dSpending}{dGroup\_Period} . \frac{Group\_Period}{Spending} = 0.3937 * Group\_Period$$

and the coefficient β is the percentage increase in Y from a unit increase in X i.e. a unit increase in a participating member after launch of social media leads to 39.37% increase in Spending.