Statistical Assignment

<Student's name>

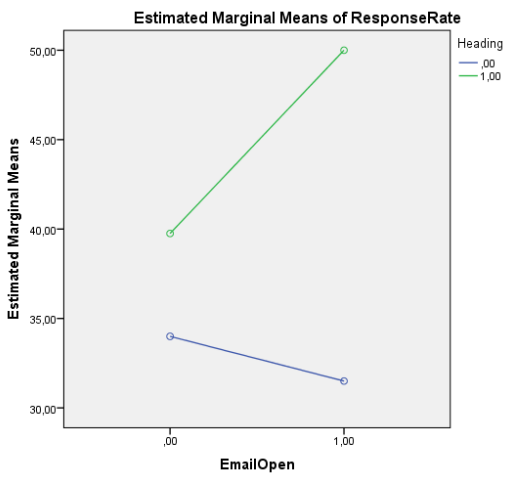
<University>

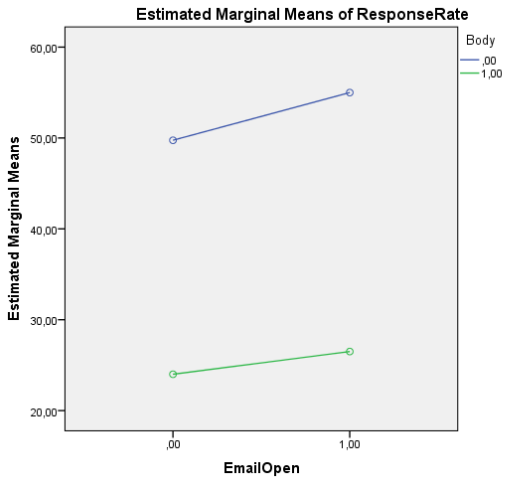
Statistical Assignment

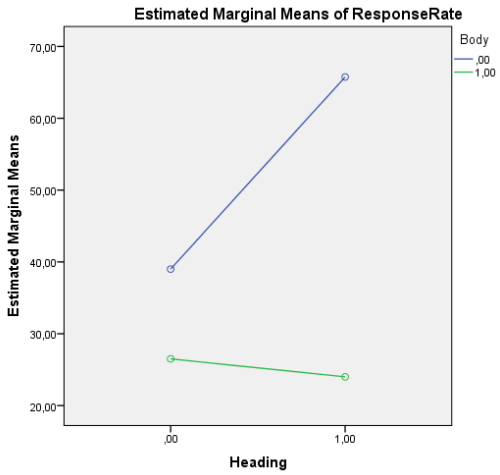
In this report, we will analyze the data collected by the company, which is interested in improving the response rate to its e-mail advertisement. The sample provided include three key factors of influence on the response rate: Heading (generic or detailed), whether e-mail was opened or not, and body of the e-mail (HTML or simple text). There were 8 different combinations established, and each combination was measured on two occasions.

In order to test cause-and-effect relationship between response rate and all three key factors, a multivariate regression analysis and repeated measures ANOVA were used. The repeated measures ANOVA is conducted to examine the interaction plots of the independent variables. A multivariate regression analysis will be conducted to find out the most significant factors of influence. The specification of the regression equation will include all three independent factors, as well as all possible interactions of these factors.

It was decided to create Interaction Effect Charts for response rate under various combinations of heading, email opens and body of e-mails. The charts were produced in SPSS and the results are provided below:







The plots above show that we should not expect a significant interaction between email open and e-mail body. However, there might be some significant interactions between heading and body, and between heading and e-mail open. The original multivariate regression equation was created on the basis of the data set, which was organized in the following way:

Run	Heading x1	Email Open x2	Body x3	x1x2	x1x3	x2x3	x1x2x3	Response Rate
1	0	0	0	0	0	0	0	46
2	1	0	0	0	0	0	0	34
3	0	1	0	0	0	0	0	56
4	1	1	0	1	0	0	0	68
5	0	0	1	0	0	0	0	25
6	1	0	1	0	1	0	0	22
7	0	1	1	0	0	1	0	21
8	1	1	1	1	1	1	1	19
1	0	0	0	0	0	0	0	38
2	1	0	0	0	0	0	0	38
3	0	1	0	0	0	0	0	59
4	1	1	0	1	0	0	0	80
5	0	0	1	0	0	0	0	27
6	1	0	1	0	1	0	0	32
7	0	1	1	0	0	1	0	23
8	1	1	1	1	1	1	1	33

The results of original regression was as follows:

Regression Statistics	
Multiple R	0.962809
R-square	0.927002
Adjusted R-square	0.878336
Standard error	6.355772
Observations	16

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	4616.875	769.4792	19.04848	0.000121
Residual	9	363.5625	40.39583		
Total	15	4980.438			

	Coefficients	*Standard error*	*t-statistic*	*P-Value*	*Lower 95%*	*Upper 95%*
Y-intercept	39.5625	4.203948	9.410797	5.92E-06	30.05251	49.07249
Heading x1	-1.125	5.50426	-0.20439	0.842598	-13.5765	11.3265
Email Open x2	20.375	5.50426	3.701679	0.004908	7.923499	32.8265
Body x3	-11.125	5.50426	-2.02116	0.073983	-23.5765	1.326501
x1x2	12.75	6.355772	2.006051	0.075809	-1.62775	27.12775
x1x3	-2.75	6.355772	-0.43268	0.675432	-17.1278	11.62775
x2x3	-29.25	6.355772	-4.60212	0.001287	-43.6278	-14.8722

The regression equation is significant ($F = 19.048$, $p < 0.001$), meaning that this combination of factors explains a sufficient proportion of variance in response rates. This proportion is mentioned in the adjusted coefficient of determination – about 87.83% of the variation in response rates is explained by this model. However, there are some coefficients, which are not significant individually. Specially, the coefficients of x1 ($p = 0.843$), x3 ($p = 0.074$), x1x2 ($p = 0.076$), x1x3 ($p = 0.675$) are not significant at the 5% level of significance.

This model can be improved by conducting a backward elimination process. We start from the elimination of the most insignificant predictor, and repeat regression analysis until all remaining coefficients are statistically significant. The final model obtained by this process is presented below:

Regression Statistics	
Multiple R	0.960715
R-square	0.922974
Adjusted R-square	0.894964
Standard error	5.905506
Observations	16

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	4596.813	1149.203	32.95206	4.57E-06
Residual	11	383.625	34.875		
Total	15	4980.438			

	Coefficients	*Standard error*	*t-statistic*	*P-Value*	*Lower 95%*	*Upper 95%*
Y-intercept	39	2.952753	13.20801	4.31E-08	32.50103	45.49897
Email Open x2	21.625	4.668712	4.631898	0.000726	11.34923	31.90077
Body x3	-12.5	4.175823	-2.99342	0.012223	-21.6909	-3.30907
x1x2	10.25	4.175823	2.454606	0.031983	1.059075	19.44093
x2x3	-29.25	5.905506	-4.953	0.000434	-42.2479	-16.2521

The final model is overall significant ($F = 32.952$, $p < 0.001$) and explains about 90.50% of the variation in response rate. In addition, all coefficients of the model are individually significant. The mathematical expression of the model is as follows:

$$y = 39 + 21.625x_2 - 12.5x_3 + 10.25x_1x_2 - 29.25x_2x_3$$

This model can be interpreted as follows. When an e-mail has a generic heading, is not opened and the body of the e-mail is a simple text, then the estimated response rate will be about 39%. If an e-mail was opened, the estimated response rate increases by 21.63%. However, if there is an HTML in the body of the text, the response rate will fall by 12.5%. Moreover, if the email was opened and there is HTML, the response rate will fall even by 29.25%. In contrast, if the heading is detailed and the email is opened, then the response rate will be higher by 10.25%.

On the basis of the above, it is suggest for the company to write e-mails with detailed headings, the e-mails should be opened, and there should be no HTML inside the body (i.e., x1 = 1, x2 = 1, x3 = 0). This combination of factors will provide the highest estimated response rates.