**Homework 1: Managerial Report on Sales Pattern Across North America and Europe**

You should use the **Sales.csv** and **Sales\_test.csv** datasets for this homework. The description of columns has been provided in a sperate file called **SalesInfo.csv**. The data contains the sale units of a **European brand** in a year at its stores across Europe and North America. They ran an experimental study to find the best product design in each market. They offered 4-types of product base on two attributes: (1) **Design** and (2) **Quality**. The design has two versions: **Luxury Design** vs. **Normal Design**. Also, the quality has two conditions: **High Quality** vs. **Normal Quality**. You can assume high quality product has higher durability. The seasonal sale units of each store have been provided in Sales.csv, where you can observe the type of the product, its price, the store’s advertising efforts, and the store’s sales force experience in that season. **You can find the details of data description in SalesInfo.csv.**

Please answer the following questions in a report document. Your submission should include your report and a SAS code. The SAS code should replicate your reported results completely. Add the SAS code as appendix to the report. Please reports your results based on **2 decimal places**.

**GOOD LUCK!**

**Guideline:** When we ask to statistically test a hypothesis, you **MUST** clearly state your **null hypothesis (H0)** and the **alternative hypothesis (Ha)**. Then, run the SAS code to test on the Data. Also, you need to clearly state what is the final result of your test and conclude based on that.

1. Describe the type of each variable in terms of a qualitative variable or quantitative variable.

How many stores are there in the data? what are their frequencies in Europe and North America? (**20 points**)

1. Plot histograms for Sales Unit, Price, and Advertising Spending. Your histograms should have the best normal and kernel fitted curve. For Sales Unit, Price, and Advertising Spending use binwidth 1000, 10, 100 respectively. Which of variables does not look different from a normal distribution? (**20 points**)
2. Provides a summary statistic of **Sales Unit, Price, and Advertising Spending**. Also, find those summary statistics by seasons? Do the **Sales Unit, Price, and Advertising Spending** have different means across seasons? Here, you Must provide a hypothesis testing to determine which of Sales Unit, Price, and Advertising Spending have significantly different means across seasons? (Hint: Use ANOVA)

* Provide a hypothesis testing to determine which of Sales Unit, Price, and Advertising Spending have significantly different means across location, i.e., North America vs. Europe? (Hint: T-test)
* Find the correlation matrix of the Sales Unit, Price, and Advertising Spending? Report both Pearson and Spearman Correlation tests. How can you describe the effects of Price and Advertising Spending on Sales Unit? Now, Use **Proc sgscatter** to plot a matrix of Sales unit, price, and advertising spending. What type of relationship do you see between Sales unit and price? (Linear or Non-linear) What type of relationship do you see between Sales unit and advertising spending? (Linear or Non-linear)? Does the scatter plot of Sales Unit based on Advertising Spending confirm your result in correlation tables? If not, why?
* Now, provide a *linear* regression of Sales unit over price? (Consider **Price** as the only covariate in the regression) Is the price coefficient significant? Does price’s coefficient make sense? What does it mean?

Now, provide a *non-linear* regression of Sales unit over price and its power 2? Are the price and its power 2’s coefficients significant? Explain your result in terms of consumers’ reaction to increase of price in this market. Compare the of the above models (linear vs. non-linear). Which model fits the data better and why?

* Now, provide a *linear* regression of Sales unit over advertising spending? (Consider **Advertising Spending** as the only covariate in the regression) Is the advertising spending coefficient significant? Does advertising’s coefficient make sense? If No, why?

Now, provide a *non-linear* regression of Sales unit over advertising spending and its power 2? Are the coefficients significant? Explain your result in terms of Sales unit changes by increasing of advertising in this market. Compare the of the above models (linear vs. non-linear). Which model fits the data better and why?

* Imagine you **only** observe **Sales unit, price, and advertising spending variables**. Based on your answers in the above parts, provide a regression model’s result to explain the Sales unit as a function of price and advertising spending. Based on Cook’D statistic, determine all influential points. You need to print the influential points in your report. Now, repeat your above regression analysis ***without*** including the influential points. Does it improve your model goodness-of-fit? (Use ) If yes, how far? (Hint: follow the rule of thumb which is stated in class)

**Please do not remove any data point for the rest of questions. You need all data for remaining questions.**

* Imagine you **only** observe **Sales unit, price, and advertising spending variables**. Based on your answers in the above parts, provide a regression model’s result to explain the Sales unit as a function of price and advertising spending. Now, check the collinearity of independent variables by reporting the VIF. If the VIF >10, drop some variables to reduce collinearity in an appropriate way. Does your new model fit the data better? (Use ) If yes why? If no, why? (**80 points**)

1. Imagine you **only** observe **Sales unit, price, and location** variables(i.e., these are your covariates). You are interested in knowing who is more price sensitive? North Americans or European. Propose a regression model to find the most price sensitive group of consumers. (Hint: Use the Interaction effect). You should write your model precisely, estimate it, and explain the estimation results to identify the most price sensitive group of consumers.

Now, imagine you **only** observe **Sales unit, promotion, and location** variables (i.e., these are your covariates). You are interested in knowing (a) do consumers respond to promotional prices? and (b) who is more responsive to promotional offers? North Americans or European. Propose a regression model to answer the above questions. (Hint: Use the Interaction effect). You should write your model precisely, estimate it, and explain the estimation results to answer the above questions.

Now, are your results regrading consumers’ price sensitiveness and consumers’ responsiveness to promotional prices **CONSISTENT** with each other? (**50 points**)

1. Imagine you **only** observe **Sales unit, price, location, Design, and Quality variables**. We know the firms offer **4-types of product** based on the level of Design and Quality. Define a new categorical variable called **Product-Type** in SAS to code the 4-types of firm’s products. Define the **Product-Type** as follows:

|  |  |  |
| --- | --- | --- |
| Design =1 | Quality =1 | **Product-Type = 1 called the Premium product** |
| Design =1 | Quality =2 | **Product-Type = 2 called the Luxury product** |
| Design =2 | Quality =1 | **Product-Type = 3 called the Durable product** |
| Design =2 | Quality =2 | **Product-Type = 4 called the Basic product** |

First, Use Box Plotting figures to determine the average sales of each type of product. Then, redo your analysis based on people located in Europe vs. North America. Do you find any evidence on a different popularity of types of product among North American and Europeans? Who does prefer higher Design? Who does prefer higher Quality?

You are interested in knowing the relationship between the **Sales Unit, Price, Product-Type, and locations** variables (i.e., these are your covariates). You should write precisely a regression model, estimate it, and explain the estimation results to identify (a) which type of product is the least popular among consumers? (b) Also, you want to know do North Americans vs. Europeans have different preference for Design and Quality? In other words, is the most important attributes of product (i.e., Design vs Quality) different among North Americans vs. Europeans? Now Are your regression results **CONSISTENT** with your Box Plotting result? (**50 points**)

1. First, Use Box Plotting figures to determine the average sales of each type of product in the different season. Do you find any evidence on a different popularity of types of product among seasons? Which product is more popular in each season? This is called seasonality effect.

Imagine you **only** observe **Sales unit, price, Season, Design, and Quality variables**. Create a categorical variable **Product-Type** to code the 4-types product as in Q5. You are interested in knowing (a) How far does the Seasonality increase the sales? (b) Also, you want to know do people have any preference to buy specific type of products in different seasons? In other words, do people buy more high quality or luxury products based on the current season? You should write precisely a regression model, estimate it, and explain the estimation results to identify the above questions. Can you explain a logical, verbal story about the seasonality effects which you find in your analysis? In other words, please try to justify your result based on a logical argument. (**50 points**)

1. Use Box Plotting figures to determine Sales Unit based on Salesforce experience level. Do find any clear evidence about Salesforce experience level? Then, redo your analysis based on the 4-types of products? Does higher Salesforce experience provide any special ability to sell more products? If yes, which type of products?

Imagine you **only** observe **Sales unit, price, Salesforce experience, Design, and Quality variables**. Create a categorical variable **Product-Type** to code the 4-types product as in Q5. You are interested in knowing (a) How far does the Salesforce experience increase the sales? (b) Also, you want to know do high experience Salesforce have any expertise to sell specific type of products? In other words, do their experience help them to sell more high-ended products? You should write precisely a regression model, estimate it, and explain the estimation results to identify the above questions.

Now Are your regression results **CONSISTENT** with your Box Plotting result?

(**50 points**)

**Guideline**: For question 8-12, consider the whole dataset, i.e., all columns. Moreover, add the power two of price and advertising spending as new predictors into dataset. Precisely, your dependent variable is **Sales unit** and the independent variables are **Price, Price^2, advertising spending, advertising spending^2, Product-Type, Promotion, Salesforce experience, Location**, **and Season**. Also, do the same procedure on **Sales\_test.csv**

1. By using Proc Glmmod, you should create two new datasets out of **Sales.csv** and **Sales\_test.csv** which include all the above quantitative variables and all dummy variables of all the above qualitative variables. By using Proc Reg, find the best regression model on the above generated data out of **Sales.csv**. you must use the Mallows’ criterion for selection procedure. Now, by considering the above best regression model, record its and calculate the by using the above generated data out of **Sales\_test.csv** (Hint: You can use Proc score)

Now, by using Proc Glmselect, find the best models according to Forward, Backward, Stepwise algorithm. In all cases, (1) - you must use the Mallows’ criterion for selection procedure, (2) adding all interaction effects at the second order, and (3) in a hierarchal order, when the main effects should be included first. Meanwhile, find and record their and the of the best selected models by the above procedures (i.e., Forward, Backward, Stepwise algorithm) by using **Sales\_test.csv** as the test dataset.

Now, please rank all the above recorded best models based on either their or the of them. If your goal is to do inference about the Sales unit based on independent variables which one of the above best models will be your choice? If your goal is to do predictive analysis about the Sales unit based on independent variables which one of the above best models will be your choice? (**30 points**)

1. Now, by using Proc Glmselect, find the best models according to Forward, Backward, Stepwise algorithm. In all cases, (1) - you must use the Cross-Validation criterion for selection procedure based on creating 10-folds of **Sales.csv** dataset (please set the seed=2), (2) adding all interaction effects at the second order, and (3) in a hierarchal order, when the main effects should be included first. Meanwhile, find and record their and the of the best selected models by the above procedures (i.e., Forward, Backward, Stepwise algorithm) by using **Sales\_test.csv** as the test dataset.

Now, please rank all the above recorded best models based on either their or the of them. If your goal is to do inference about the Sales unit based on independent variables which one of the above best models will be your choice? If your goal is to do predictive analysis about the Sales unit based on independent variables which one of the above best models will be your choice? (**30 points**)

1. Now, by using Proc Glmselect, find the best models according to LASSO and Elastic Net. In all cases, (1) - you must use the Cross-Validation criterion for selection procedure based on creating 10-folds of **Sales.csv** dataset (please set the seed=2), (2) adding all interaction effects at the second order, and (3) in a hierarchal order, when the main effects should be included first. Meanwhile, find and record their and the of the best selected models by the above procedures (i.e., LASSO and Elastic Net) by using **Sales\_test.csv** as the test dataset.

If your goal is to do inference about the Sales unit based on independent variables which one of the above best models will be your choice? If your goal is to do predictive analysis about the Sales unit based on independent variables which one of the above best models will be your choice? (**30 points**)

1. Now, by using Proc Glmselect, do the regression analysis based on Bagging Method. In all cases, (1) - you must use the Cross-Validation criterion for selection procedure based on creating 10-folds of **Sales.csv** dataset (please set the seed=2), (2) adding all interaction effects at the second order, (3) in a hierarchal order, when the main effects should be included first, and use **100** samples (based on Bootstrapping). Based on the Average Model of 100 models, find and record the and the by using **Sales.csv and Sales\_test.csv** as the training and test datasets respectively (**30 points**).

1. This firm wants to increase its sales. They have two options. The first way is to train their sales-force people to become more expert. The second way is to increase their advertising budget. The firm told you that they can increase their advertising budget by **10%**. Also, they know, by having a one-day workshop for “Low” and “Average”-types of sales-force people, they will become “Average” and “High”-types sales-force respectively.

**You want to do a contrafactual analysis for the firm**. *You will do the analysis here based on the Average Model of 100 models in Q11*. Replicate the above scenarios by *either increasing the advertising spending variables by 10% in* ***Sales.csv***, *or increasing the sales-force experience level form “Low” TO “Average” and from “Average” To “High” in* ***Sales.csv*** *(due to a one-day workshop)*. Based on these changes in **Sales.csv**, you will have two new datasets called **Sales\_increasing\_ad and Sales\_increasing\_expertise**. Now, predict the sales unit in the above two datasets based on the Average Model of 100 models in Q11. You know the difference of sales before and after of the above firms’ strategies (i.e., either increasing advertising budget by 10% or increasing sales-force expertise by holding workshop). Which strategy will increase the sales more? What is the average sales unit changes by Product-Type in each strategy? Which strategy is more profitable from your point of view based on your information about this firm? (**100 points**)

1. Provide only one-page which summarizes your analyses in this homework (or any other important trends you find). For example: Which variables have positive effects on the Sales Unit? Which variables have negative effects on the Sales Unit? Does exist a systematic (i.e., significant) difference between Europeans and Americans? If Yes, what are they? Who is more price sensitive? Which product should be targeted and to whom? What should be the strategy on Salesforce training to increase the sales? And … (**100 points**)

**Guideline**: For question 13, you should provide a professional, managerial one-page to summarize your analyses. It must be short, informative, and inclusive. ***The last sentence must provide a clear strategy for this firm***. Which product should be targeted in each market & what should be the price level in each market? What can be done to increase the sales? Question 13 will be marked very competitively. The best answer will receive the full mark. We will decrease marks according to your rank among groups.