Name: **­Shirazi Siddiqui**  Date: **25th Feb 2020**

* **Clearly indicate your answer**
* **This is a Quiz - you must work alone!**
* **Open book and notes.**
* **Each problem is worth 9 points.**
* **Please submit the completed Quiz before the next class time.**
  + **If late in turning your quiz in, you will automatically lose 30pts**

**Dataset also contains the Score Data (2nd tab).**

**Problem 1**

Data set you have given does not have missing data or outlier. Is there any variable skewed enough to warrant log transformation?

**Descriptive Statistics of Data using Excel:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Taste** |  | **Nutrition** |  | **Marketing** |  | **Price** |  |
|  |  |  |  |  |  |  |  |
| Mean | 6.71875 | Mean | 5.929688 | Mean | 5.648438 | Mean | 6.15625 |
| Standard Error | 0.189843459 | Standard Error | 0.180967 | Standard Error | 0.143164 | Standard Error | 0.23172302 |
| Median | 7 | Median | 6.5 | Median | 5 | Median | 6 |
| Mode | 8 | Mode | 7 | Mode | 5 | Mode | 6 |
| Standard Deviation | 2.147833554 | Standard Deviation | 2.047405 | Standard Deviation | 1.619711 | Standard Deviation | 2.62164671 |
| Sample Variance | 4.613188976 | Sample Variance | 4.191868 | Sample Variance | 2.623462 | Sample Variance | 6.8730315 |
| Kurtosis | -0.285865656 | Kurtosis | -0.49523 | Kurtosis | -0.47926 | Kurtosis | -1.01579654 |
| Skewness | **-0.523791291** | Skewness | **-0.3175** | Skewness | **0.463012** | Skewness | **-0.17297612** |
| Range | 8 | Range | 9 | Range | 7 | Range | 8 |
| Minimum | 2 | Minimum | 1 | Minimum | 2 | Minimum | 2 |
| Maximum | 10 | Maximum | 10 | Maximum | 9 | Maximum | 10 |
| Sum | 860 | Sum | 759 | Sum | 723 | Sum | 788 |
| Count | 128 | Count | 128 | Count | 128 | Count | 128 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Taste (Skew)** | **Nutrition (Skew)** | **Marketing (Skew)** | **Price (Skew)** |
| **-0.523791291** | **-0.317500649** | **0.463011759** | **-0.1729761** |

Skewness calculated using formula “**skew**” in excel :

Both methods giving same coefficient of skewness of variables : Taste, Nutrition, Marketing and Price.

If Coefficient of skewness is < -1 or > +1 then it is skewed enough for Data Transformation.

The Coefficient of skewness is within the range of **-1< coefficient of skew < +1** for Taste, Nutrition, Marketing and Price. **Hence, Data Transformation is not needed in the model.**

**Problem 2**

Bring the data in to the SAS Enterprise Miner. **Please have Regression, Neural Nets, Decision Tree, PLS and LARS for your predictive models.** Which model is the best based on Validation Data – Average Squared Error? Please copy and paste the comparison node result.

**Master Predictive Flow:**

**A close up of a map

Description automatically generated**

**Model Comparison Result:**

A screenshot of a cell phone

Description automatically generated

**Auto Neural, Multiple Regression and PLS** are the best models as **Averaged Squared error** is smallest for these cases i.e. **0.00229**

**Problem 3**

Using the attached Score Data, based on the best model selected in Problem1, what is the predicted share for ID?

1. **Base Case: 24.75%**
2. **Optimistic: 30.03%---Highest as expected**
3. **Pessimistic: 17.83%**

**Predictive Model Flow with score data:**

**A close up of a map

Description automatically generated**

**Problem 4**

Please paste your screen of the predicted scores here.

A screenshot of a cell phone

Description automatically generated

**Problem 5**

Please do a sanity (common sense) check of the predicted scores. Explain why it does or does not make sense.

**Answer:**

It does make sense as best model giving the best result : Optimistic approach is better than Base case as well as pessimistic approach. Optimistic approach predicts 30.03% of Market share. Base case predicts 24.75% of Market share and Pessimistic approach predicts 17.83% of Market share.

1. **Base Case: 24.75%**
2. **Optimistic: 30.03%---Highest as expected**
3. **Pessimistic: 17.83%**

**Problem 6**

If you have a situation where there are more input variables than the number of observations, what can you expect from your develop predictive model?

1. **You’ll have a model that overfits to the training data**
2. You’ll expect to use only 1% of the input variables for the model
3. Your model will perform nearly well on the validation data as it did with the training data
4. Your neural networks model is expected to be the best choice in such situation
5. None of the above

**Problem 7**

Explain a situation where you think is most suited for PLS-based and LARS-based predictive model?

**Answer:**

PLS regression is most suitable for the scenarios where **inputs variables are too large** as compared to number of observations. It is also used when redundant variables are prevalent (multiple collinearity) and where Input variables do not have well understood relationship with target.

LARS is a variation of a regression method that is similar to stepwise regression and quite stable method. It is an Efficient algorithm which works well with **correlated inputs**. It can be **cross validated to tune the model**.

**Problem 8**

If you want to predict accurate house price in Pennsylvania, what variables will you collect, why would you collect these variables, and how would you go about collecting it to minimize bias?

**Answer:**

**Below variables are required to collect in order to predict the price of the house:**

* Zip code (Location)
* Area of House (sqft\_above/sqft\_basement/sqft\_living area/sqft\_lot/garage)
* How old is the house?(year of built, year of renovated)
* If security is available
* How close/far is the market
* Connectivity of house location with transport
* How many floors does the house have?
* What material is used in the construction
* Weather Condition (Pool/Fireplace)
* If terrace/Patio is available.
* If car parking is available
* Play area / parks for kids (if any)

**Reasons to collect above variables:**

* A model can be built with numerous variables that are tested for impact on the value of a property, such as square footage and the number of bedrooms etc.
* There is always the danger that variables contain autocorrelation and/or multicollinearity, or that correlation between variables is spurious.

**Right variable collection to avoid bias:**

Traditional valuation methods all have a significant risk of bias.

1. Ideally, variables within a model won’t be correlated to each other.
2. The purpose of model building is to find useful variables that will make valid predictions

Variable selection can be one of the most difficult parts of the process without statistical

software. However, SPSS allows us to quickly build many models from a combination of

variables we’ve deemed appropriate for a linear regression. SPSS will automatically [filter](https://www.spss-tutorials.com/stepwise-regression-in-spss-example/)

[out variables](https://www.spss-tutorials.com/stepwise-regression-in-spss-example/) based on our thresholds for statistical significance and return only the best

models.

1. There is also an emphasis on predicting variables, such as the rate of return in the income approach. Eliminating the need for this prediction could be attractive to many real estate investors.
2. There should be a linear relationship between the inputs (independent variables) and the output (dependent variable). For example, we could assume that there is a linear relationship between the heated square feet in a home and its overall value. However, due to diminishing returns, we could discover that the relationship is non-linear, requiring a transformation of the raw data.
3. The independent variables should not be random. Put simply, the observations for each independent variable in the model are fixed and assumed to have no error in their measurement. For example, if we are using the number of units to model the value of an apartment building, all of the buildings in our sample data will have a fixed number of units that won’t change, regardless of how we build the model.

**Problem 9**

A real estate firm in California wants to use the model developed in Problem 8 for California’s housing market without any modification. Would it work – why or why not?

**Answer:**

It will not work as location matters a lot when it comes to price of a house. There is a big difference in the cost of living of Pennsylvania and California.

Traditionally, there are three approaches for valuing property: comparable sales, income, and cost.

- Overall, Philadelphia, Pennsylvania is 41.6% cheaper than Los Angeles, California  
- Median Home Cost is the biggest factor in the cost of living difference.  
- Median Home Cost is 77% cheaper in Philadelphia.

A salary of $113,000 in Los Angeles, California could decrease to $64,174 in Philadelphia, Pennsylvania.

**Problem 10**

Please state which predictive model you like the best out of all the methods learned in the class, and why. And using this model, run the predictive model to get the results for the following IDs. Compare the result with that of Problem3. If your model of choice is the same as that of Problem3, please choose the 2nd best model (shown in model comparisons) for the comparison purposes.

I like multiple regression model as the best model because there are two main advantages to analyzing data using a multiple regression model. The first is the ability to determine the relative influence of one or more predictor variables to the criterion value. The real estate agent could find that the size of the homes and the number of bedrooms have a strong correlation to the price of a home, while the proximity to schools has no correlation at all, or even a negative correlation if it is primarily a retirement community.

The second advantage is the ability to identify outliers, or anomalies. For example, while reviewing the data related to management salaries, the HR manager could find that the number of hours worked, the department size and its budget all had a strong correlation to salaries, while seniority did not. Alternatively, it could be that all of the listed predictor values were correlated to each of the salaries being examined, except for one manager who was being overpaid compared to the others.

**2nd best model is LARS : Connected Score with LARS and rerun the score.**

A close up of a map

Description automatically generated

A screenshot of a cell phone

Description automatically generated

1. **Base Case 2: 21.60%**
2. **Optimistic 2: 27.82% ---Highest as expected**
3. **Pessimistic 2: 19.93%**

It does make sense as 2nd best model giving the 2nd best result : Optimistic approach is better than Base case as well as pessimistic approach. Optimistic approach predicts 27.82% of Market share. Base case predicts 21.60% of Market share and Pessimistic approach predicts 19.93% of Market share.

**Problem 11**

Please read the following article, and briefly summarize key points.

<http://edge.org/conversation/richard_nisbett-the-crusade-against-multiple-regression-analysis>

**THE CRUSADE AGAINST MULTIPLE REGRESSION ANALYSIS**

Richard was more interested in things that become cause against correlational statistical analysis (**Multiple Regression analysis**). For instance, how Vitamin E is associated with lower prostate cancer risk. Look at the correlational evidence and indeed it turns out whoever take Vitamin E have lower risk for prostate cancer. The correlational and observational evidence tells a story and the experimental evidence tells something completely different.

1. In health data, the big problem is the healthy user bias because who’s taking Vitamin E also doing rest all right things like taking Vitamin E, watching weight & cholesterol, doing exercises, drinks alcohol in moderation, doesn’t smoke, has a high level of education and a high income. All of these things are likely to make you live longer. That’s all correlated and looks like Vitamin E is terrific
2. The New York Times recently published an article that people who have elaborate weddings tend to have marriages that last longer and who makes elaborate plans for expensive weddings: people who are financially stronger, people who are more educated, people who are richer, people who are older, the later you get married, the more likelihood that the marriage will last. No, it’s because those people are the types who have a good prognosis for a lengthy marriage.

Health statistics in general should be extremely uncertain about this experimental study. The consequences of this junk research are enormous. Here Richard is trying to find ways to get people to stop doing it and to make the general reader aware that they have to ask themselves, "**Do I think that this is a correlational study or is it an actual experiment**?"

He pointed out how weak our predictions can be:

The predictability to a correlation of about 0.3, which is not a very strong relationship. At the level of one situation to another, how extroverted is Bill in Situation A and B that correlation runs about 0.1, which means it’s about a 54% chance that would get it right as to whether the more extroverted person in Situation 1 would be the more extroverted person in Situation 2. That's up from the 50%. When the person’s attention is directed slightly differently, you may not get the effect.

We have dozens of experiments that make the same general point, in this case, you’ve been telling lies or truth depending on what color light is on. When the experiment is over and you’re asked to express an attitude and one of those lights is on that tend not to be as confident about what you’re saying if it was the "lie light" that was on. That’s doesn’t replicate but the general point is that some minimal seeming cue, which gets into the stream of behavior and has an impact. We have hundreds of those studies, many individual ones that are not highly replicable.

There’s tremendous redundancy in what we’re telling the public about human behavior. Though it may not be the case that you can tell a bunch of college students words that remind them of old folks and they’ll be walking more slowly. There are many studies making a point similar to that.

Let’s have an example, you put three dots over the coffee urn and the dots are in the position two up, one down, which is like more of the human face: two eyes and a mouth. You watch to see how many people leave money in the honest box when those dots are there. People leave more money in the honest box when the dots that are meaningful of a face looking at them are there. If you invert it, two dots on the bottom and one dot on top, it has no effect.

If you have loud music playing and many people are in the room, that might not have any effect because people don’t notice it or it has a minimal effect. If that failed to replicate, it wouldn’t be surprising and it wouldn’t shake your faith in the idea that minimal cues, which ought to have no part in determining your behavior, do have an impact. We are constantly being influenced by things that we don’t recognize have had an influence, and which are sometimes embarrassing to know. That isn’t why we’re unaware of them. We’re unaware of them because we don’t have access to our cognitive process. We claim that we do.

You ask me why I do something, I’ll give you an answer, although you’ll probably believe it more than I will because I’m so aware of the extent to which we’re unaware of what goes on.

First experiment was a learning experiment; people memorize word pairs. There might be a pair like "ocean-moon". Then we’re through with that experiment: "Thank you very much." Other experiment on free association, where you are to give an example of something in a category we give you. So one of the categories is "Name a laundry detergent." People are much more likely to mention Tide, if they’ve seen that ocean-moon pair.

Two of the things that have gotten some notice and that were an attempt to bottle a personal experience:

1. When he was in college, he had bad insomnia. Just couldn’t get to sleep. What would happen is that lie in bed and start worrying about my relationships, or about the exam, and then get more and more worked up about it. Eventually he was hot and throwing off all the covers, tossing and turning. Very early in his career he wondered if we couldn’t break this escalating pattern of worrying about something. The arousal you experience is evidence of how worked up you are about this thing. If we gave insomniacs a pill to take at bedtime and said, "We’re interested in people’s dream content and want to see how it’ll be affected by this pill which will cause your heart to pound a little more irregularly, a little faster, breathing to become more shallow and more rapid. You might find yourself getting warm, sweaty palms etc. People who take a pill and given those instructions get to sleep more quickly. It turns out that you can interrupt the vicious cycle by getting people to attribute the arousal to something non-emotional.
2. He didn’t get a terribly good K12 education. He grew up in El Paso, Texas and always regarded this as a strength for me because I didn’t have to do a lot of homework, so I would read books. I would sit in class and fantasize about the books I was going to read. In college I took an American literature course. "This is great. I’ll get credit for stuff that I would do anyway." After a while I’m reading these books, early American literature. The question then becomes, can you get people to do something in a mood where they’re thinking that it’s essentially work, it’s something they’re doing in order to get something else, versus something that they’re doing as an expressive thing? They’re doing it because it’s fun.

We put some “**Magic Markers**” on a table that the kids had not seen before. We watched how much each kid played with it. After a week, a nice man comes up to each kids and says, "Do you remember the Magic Markers that were here a week ago? I would like to see what kinds of pictures you would draw, and if you would be willing to draw some pictures for me with these pens, I could give you “**Good Player Award**”. It has a place for your name, a gold seal, and a blue ribbon. Would you like to have a chance to get that?" All of them do, of course. Or we don’t ask them to draw or offer them anything.

We come back a couple of weeks. The Magic Markers are out again. They haven’t been there during the two weeks. We’d look at how much kids draw with it. The ones who have contracted or made a bargain that they can get this other thing if they play with it, play half as much as the kids who are not offered this contract. It isn’t something about the certificate, because some kids we didn’t make a contract with are great. I want to give you this award. Those kids were just like the control kids. It wasn't that there’s something bad about getting the award; it’s that something’s bad about contracting, about framing the thing as work.

In World Economic Forum meeting, He was on a panel of people who were to think of ways to make people behave in their own interests and in society’s interests. There were economists, psychologists, political scientists etc. One word kept coming up over and over again: incentivize. You incentivize, and usually it was money. He hears the word incentivize and says, ***"If imagination fails, incentivize."***