STONY BROOK UNIVERSITY, APPLIED MATH AND STATISTICS

AMS572 Group Project
Questions & Answer

Group 1

November 30, 2017

1 QUESTION ONE

- **Question:** One of the assumptions that needs to be verified when anova test is implemented is normality. Did you check the variables for normality?
- **Answer:** Each data point is independent. Number of datapoints > 30. Hence the data is assumed to follow normality assumption. We spoke with the Professor and because of the scope of the class and the methods we went over to analyze data, we decided it was acceptable to make an assumption of normality for this dataset. Our sample size of 30,000 points was also sufficient to apply CLT. However, we also demonstrated a boxplot graph and individual pair-wise comparison in addition to the ANOVA tests.

2 QUESTION TWO

- **Question:** Hypothesis 2 focuses on delivering a logistic regression model that will predict whether a person will default or not. One can add many independent variables that show so be statistically significant but intuitively have no meaningful relationship. Th weights assigned to all the different outstanding bills have different signs.Is meaningful to assign outstanding bills for specific months (for example May, June, July) as separate independent variables and if it is justify it?
- **Answer:** There are many variables in our dataset: age, gender, education, history of past payment, amount of bill statement, amount of previous payment and so on. Among

them, the payment behaviors are the most important factors when the banks make decisions about whether there is a payment defaulter. In our dataset, the history of past payment, amount of bill statement and amount of previous payment are the indicators related to the payment behaviors of people. So, firstly, we chose these variables as candidates. Next, we also computed the correlation between these variables and found some correlation coefficients between certain variables are indeed over 0.7.

But why we finally didn't delete the variables with high correlation? In linear regression, like our hypothesis 3, if certain variables have high correlations, some of them have to be deleted in order to avoid multicollinearity, although some real information will be lost. But multicollinearity is more related to the accuracy of the coefficients in the model, instead of the result. In hypothesis 3, coefficients are just what we want. But in logistic regression, we care more about the result. So, we chose to keep these variables to avoid loss of the real information. Particularly, in our dataset, the amount of bill statement and the amount of previous payment must have relationships because they have some autocorrelations actually. It's also related to time series analysis and beyond our class.

If we delete the variables with high correlations, we can get almost similar prediction accuracy (I have already tried but didn't show the results in the slides). However, it's very hard to explain the reason in reality. For example, if the correlation between second month and the third month is high and between the fourth month and the third month is low, I have to find the real reasons of people's behaviors actually. Plus, if I choose the dataset during another time period with same variables, the correlation in the example may be reverse. So, I think keep the variables I have chosen is a good choice.

There is a also famous application in logistic regression, which is about MNIST handwritten digit database. The accuracy of logistic regression model in this dataset can be more than 90%. (https://github.com/harshkn/MNIST_Logistic_Regression) Actually, I think, one pixel and the pixels around it can be highly related. But when it comes to prediction, it still get excellent results.

Therefore, just focus on some months won't change the accuracy a lot but it will be hard to explain the real meanings and the model will become also limited. So I think our model is a more general model with suitable number of indicators and good accuracy.

3 Question Three

- **Question:** We do see log likelihood graph. How were the multiple linear regression coefficients estimated? Was it using OLS or Maximum likelihood ?

- **Answer:** It's related to the Box-Cox transformation which we used in our third hypothesis. Box-Cox transformation is a method we usually use, if we find that the distribution of the residuals doesn't follow a normal distribution.

  Assume Y is a random variable. The Box-Cox transformation of Y has the form:

$$Y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ log(y) & \lambda = 0 \end{cases}$$

  The basic idea for the Box-Cox transformation is that we want $Y^{(\lambda)}|\mathbf{X}$ follows a normal distribution. But how to choose the best lambda for this transformation? Here we use the Maximum likelihood method.

  Assume that there exists a lambda, which satisfied the linear model:

$$Y^{\lambda} = \mathbf{X}\beta + \epsilon$$

  where $\epsilon \sim N(0, \sigma^2)$. So when given $\mathbf{X}$, $Y^{(\lambda)}|\mathbf{X}$ should follow the normal distribution $N(\mathbf{X}\beta, \sigma^2)$.

  Then the joint density function for $Y^{(\lambda)} = (y_1^{(\lambda)}, y_2^{(\lambda)}, ..., y_n^{(\lambda)})'$ should be

$$f(Y^{(\lambda)}) = \frac{1}{(2\pi\sigma^2)^{2/n}} \exp\{\frac{1}{2\sigma^2} \| Y^{(\lambda)} - \mathbf{X}\beta \|^2\}$$

  So the likelihood function for $Y = (y_1, y_2, ..., y_n)$ is

$$L(\lambda, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{2/n}} \exp\{\frac{1}{2\sigma^2} \| Y^{(\lambda)} - \mathbf{X}\beta \|^2\} * \prod_{i=1}^{n} y_i^{\lambda-1}$$

  And the log-likelihood function would be

$$\begin{aligned} l(\lambda, \beta, \sigma^2) &= \log L(\lambda, \beta, \sigma^2) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \| Y^{(\lambda)} - \mathbf{X}\beta \|^2 + (\lambda - 1) \sum_{i=1}^{n} log(y_i) \end{aligned}$$

  As we know form this course, for given λ, the maximum likelihood estimation for β and σ^2 is

$$\hat{\beta} = (X'X)^1 X' Y^{(\lambda)}$$

$$\hat{\sigma^2} = \frac{\| Y^{(\lambda)} - \mathbf{X}\hat{\beta} \|^2}{n}$$

  So, for given λ, the maximum likelihood function is

$$l_p(\lambda) = l(\lambda, \hat{\beta}, \hat{\sigma^2}) \quad = C - \frac{n}{2}\log RSS(\lambda) + (\lambda - 1) \sum_{i=1}^{n} log(y_i)$$

  The log-likelihood graph you mentioned is just the plot of this function. And the best lambda is

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, l_p(\lambda)$$

  Here we just calculated values of the function with different λ's to choose the best λ.