BUDA 530 Final— Spring 2020

Group 5

2/19/2020

### Problem 1 (20 points)

In your own words address the following questions about maximum likelihood estimation. What is maximum likelihood estimation? What is the procedure for finding the maximum likelihood estimates? Is there a relationship between the ordinary least squares estimates and generalized linear models that is based on maximum likelihood estimation? If so give insights on that relationship.

Simply stated, the maximum likelihood estimation is an attempt to use a standard distribution model to best fit a set of known data points. We are trying to estimate the parameter data given the data we know. It really is the case of making the curve fit the points in the hope that it will most likely fit all the points including those not yet measured. Once the model fits properly, we may use it for either inference or prediction. The models used are distributions and could be a normal, binomial, exponential or something else. Using a distribution function is important because if the model is a known type of distribution (with understandable properties and behaviors), then it will be easier to work with and may be reused for similar experiments.  
In addition to the type of distribution, we also need to consider the type of parameter to be solved for. In a normal distribution, we would solve for mean and standard deviation. In a binomial model, it might be the “p” probability of success for the trial. Whatever the parameter, solving it will help us identify where the curve should fit in order to most optimally match the available data. We use probability density functions (PDF) for continuous variables and probability mass functions (PMF) for discrete variables.

Regardless of the type of distribution, we know that the curve will either have a maximum or minimum value. Identifying this is how we fit the curve to the data. So, we will need to determine the value of that maximum (or minimum). This is done by taking the derivative of the distribution to identify where the slope = 0.  
We then take the derivative for the parameters we’re trying to solve. For example, in a normal distribution this could mean searching for both the mean (μ) and standard deviation (σ). In this case, the user would need to take derivatives of the distribution function first for the mean and then again for the standard deviation.  
The problem is that the formulas for distributions can be quite nasty to work with. Taking the derivative of a normal distribution can be complex. However, calculus allows us to take log of the function which makes taking the derivative a little easier. Using the log likelihood simplifies the derivation allowing us to treat nonparametric terms as constants. It may not be quick, but it will be quicker than taking the derivative of the original equation. Caution should be taken when trying to interpret the coefficients in a log likelihood model. The results may need to be transformed back before being interpreted.

Ordinary Least Squares (OLS), Generalized Linear Models (GLM) and maximum likelihood estimation are related. OLS and maximum likelihood are both used for fitting data using linear modeling. The equation for OLS in fact comes from a similar process of finding the maximum likelihood estimation for a GLM. Both processes take the first derivative of a complex equation and then solve for the parameters. So, for a normal distribution the OLS parameters would be identical to those found by maximum likelihood estimation.

### Problem 2 (20 points)

Here are some common problem occurring in industry right now. Part of your education is to be able to analyze data, but also start to think about the type of analysis needed to answer the question. Most of your clients will never see what happens behind the scenes but they will be impacted by the results. As much as we would like it, most stakeholders don’t know the analysis needed they just know the question they need an answer too. Below I’m going to describe 2 scenarios. Suggest the type of analysis/model you would used based on the methods you learned in this course. Give justification for your selection.

A company that offers lines of credit (think bank or credit cards) is interested in the risk associated with their customers. There are many different ways to view this, but they plan to use this information to offer better rates, or make offers to new customers based on information they have on these customers. The information available is income, employment history, home ownership, “some” credit history, and current credit liability for a credit report. We have credit score, but are looking to supplement it with our own metric and recommendations. We believe that using credit score only may be too restrictive to customers who will pay appropriately. We are actively looking at the risk associated with these lines of credit. For right now these loans are for a fixed credit limit.

**Scenario 1** One stakeholder is interested in different levels of risk of default/non-payment. They have provided information on 5 levels of risk, of every individual for their opinion and would like you to create a model based on this. The 5 levels of risk, are “extremely low risk”, “low risk”, “moderate risk”, “high risk”, and “extremely high risk”. They ask you to find relevant predictors for this metric and report a model they can use in an automated procedure to provide these recommendations on top of what they already have. Recommend a statistical model that you can use and how you would evaluate this model relative to customer needs. Are there issues with the method you’ve proposed? How can you explain this to your client?

The first thing we identify here is that we would use a multinomial regression model as the stakeholder has given us 5 levels of risk they assign to customers. Multinomial logistic regression is used to model nominal outcome variables, in which the log odds of the outcomes are modeled as a linear combination of the predictor variables. The model is ordinal in nature, meaning there is a natural order that occurs through the distribution of default risk levels, from low to high risk. The risk levels we would use for the response of our model are extremely low risk, low risk, moderate risk, high risk, and extremely high risk. Our beginning set of predictor variables would be income, employment history, home ownership, “some” credit history, and current credit liability as made available by the company. We can fit the multinomial with all the predictors and then look at the coefficients from the summary. From the coefficients we will be able to see the log ratio of probabilities increases and decreases for each risk level compared to the baseline category (the baseline category would depend on how we code the fit model). After we look at the coefficients, we can run test and predict functions on the fully fit model to see how many correctly identified customer risk levels our model had. If the fully fit model doesn’t give us a high accuracy, we can then investigate if some predictors are not needed and fit multiple other models. We can incorporate the step function to see which predictors can be removed and use AIC to compare multiple models.

One issue that can arise after we have fit our model is that if the stakeholders change their interpretation criteria for their risk levels, we will need to go back and change the model. The model we fit would only be valid with criteria staying constant. Another issue could be as individuals were giving risk levels to customers based on their opinion, some employees may have been looking at different standards to set this level. This could create some inconsistencies across the data that would alter our model results.

However, this model’s positives outweigh potential issues to explain to our client. We would be able to talk about which predictors are increasing risk decisions the most (won’t get too deep into explaining log ratio probabilities of coefficients), we can show how different models predicted based on certain combinations of predictors, and also show prediction probabilities at the individual level. All these would give our client great insight on the data they have regarding assigning customer risk levels.

**Scenario 2** Another stakeholder is interested in only defaults. Since credit is offered to those who the company deems appropriate to offer credit to there is a certain “filter” on who will default. Basically defaults become a “rare” occurrence, but we still want to be able to predict. The data the company provides shows a 15% default rate, which they deem acceptable, but they want to see if they can identify contributory factors to defaults. Prediction at an individual level is done through another means. They want you to develop a statistical model that can identify the number of defaults that can occur in a specific month. They also suggest there could be a time effect (month may matter) and number of defaults the previous month may matter as well. Explain the model you would try to implement and why it makes sense. How would you evaluate this model and method? What are the issues that can arise with your approach? How would you explain this to your client?

The stakeholder is looking for a statistical model that can identify the number of defaults that can occur in a specific month. There are a few ways that we could approach this request but using a time series forecast would be the best choice. Since the company already has a prediction model for defaults at an individual level, we want to examine the total number of defaults for future months. First, we can fit a time series model to get a visual of how the total number of defaults has moved over time. Then since the client mentions that there could be a time effect to suggest defaults, we would want to explore a Holt Winters type model or use an ARIMA model depending on the seasonality of the data. Once we fit our best model, we can predict off the model for future months to relay back to our client.

We see a couple issues that could arise from doing this type of modeling. One problem could be if we do not find seasonality in the data, we would get wide confidence intervals that do not give much useful insight to our client. Another issue could be that since defaults are considered a rare occurrence, our forecasting could be limited. This would also give us wide margins for prediction intervals that would not serve any purpose to our client. The amount of data the client gives us can also be an issue. They do not specify how much historical data we are getting for us to forecast off. If we are getting many years’ worth of data, our forecasts would be more accurate than only getting one or two years. Also forecasting models are usually best served for predicting 1 or 2 months in the future. Once we pass that point our intervals will tend to get larger.

Well created forecasting models are something that can be easily reviewed with our clients. Being able to show graphically our findings would be something they can understand quickly. If we can show seasonality and trends over a certain time of the year, our clients would be better prepared and equipped the next time around for a spike in defaults.

### Problem 3 (20 points)

In this course we’ve discussed generalized linear models, time series analysis, and some subsets of addative/non-linear models. In your own words and in less than a page describe how these three concepts are relate to one another? If there is no relationship between these methods explain why they are not connected?

Generalized linear models (GLM) start to look at responses that are not continuous. Generalized linear models give us a general frame work and structure (data, linear predictors, and mean-variance relationship) in which we can use to model and predict a variable (y) for a set of given predictors (x). Generalized linear models allows response variables that have arbitrary distributions, and for a link function of the response variable to vary linearly with the predicted values. The mean, of the distribution depends on the independent variables. Depending on the mean variance relationship will determine which GLM framework we will use whether it is normal, logistic, or poisson.

Additive/non-linear models involve creating different types of transformations on the data. At times creating localized predictor variables to leverage trends over a given time. Additive/non-linear models give us a set of tools to deal with non-linear trends with the factors (x) and can be incorporated into generalized linear models to help improve predictions. In a non-linear model, the structure includes determining the proper model to use and requires starting values before transformation can occur. These values are typically in the form of minimums or maximums. Non-linear model, do not require the relationship between variables nor does the data have to be linear in nature. Unlike linear regression and time series analysis, non-linear models can have more than one parameter per predictor variable. The downside of using an additive and non -linear models is that we lose interpretability but we can improve our prediction accuracy in certain localized cases. By leveraging transformations like polynomial regression, step and basis functions, and regression and smoothing splines we can manipulate certain Y and X relationships to improve our overall prediction accuracy as not all relationships may be linear.

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. Time series analysis involves predicting outcomes from outside the bounds of our data. Time series analysis looks at trends for a given Y with respect to a time component (x). This is often used for forecasting metrics to predict future metrics that is based on historic trends. We are often using this to predict counts or other metrics around “number of x”. In a time series model, the data must be defined at fixed intervals and typically in a specific order prior to conducting analysis of any kind.

Generalized linear models, non-linear models, and time series models are forms of regression analysis that are related in that all three are developed to predict the probability of an event that occurred, as related to a dependent variable, was indeed statistically significant. All three models depend on a defined data structure to enable the possibility of fitting the output in the most appropriate way to be impactful.

Time-series models usually forecast what comes next in the series - much like pattern. While in regression models, regression can be applied to non-ordered series where a response variable is dependent on values taken by predictors. When making a prediction, new values of variables are provided to evaluate the response. GLMs will have arbitrary distributions while time series will have more of same pattern. GLMs are rather to model variable Y as function of some other variable. While in the time series models you are mostly modeling variable Y as function of itself, but from previous time steps.

### Problem 4 (20 points)

In this problem we want to study the labor participation of women in the 1980’s. The Mroz data in the car package depicts labor force participation. Look at the help file for definition of the variables. We are going to use lfp as our response in this case. Create and decide on a model (you need to fit a few), that models this, discuss relevant variables and the impact they have. You can use lwg in your model as it depicts what income would be expected if that individual worked. At least one of your models must, use splines or non-linear terms.

After looking at the data it was determined that we would be fitting a binomial logistic regression model since we have a binary response (yes or no). First we fit a full binomial logistic regression model then leveraged the step function to reduce the number of features.

After running the full model through the step function we saw the model reduced and the variables k618 (number of children 6 to 18 years old) and hc (husbands college attendance) removed and an AIC improvement from 921.27 to 918.46. Next we decided to determine if there were possible non-linear terms that we could leverage transformations on to improve the model further.

AIC(m1,m2, m3, m4)

## df AIC  
## m1 6 918.4554  
## m2 7 765.7259  
## m3 14 769.5362  
## m4 8 767.8191

anova(m1,m2, m3, m4, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: lfp ~ k5 + age + wc + lwg + inc  
## Model 2: lfp ~ k5 + age + poly(lwg, 2, raw = FALSE) + poly(inc, 2, raw = FALSE)  
## Model 3: lfp ~ k5 + hc + age + bs(lwg, knots = 1) + bs(inc, knots = c(25,   
## 50, 75))  
## Model 4: lfp ~ k5 + age + hc + k618 + wc + s(lwg) + s(inc)  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 747 906.46   
## 2 746 751.73 1.000e+00 154.730 <2e-16 \*\*\*  
## 3 739 741.54 7.000e+00 10.190 0.1781   
## 4 739 739.82 -3.109e-05 1.717   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We then compared 4 models. The first model was our binomial logistic regression, the second model had polynomial transformations on lwg (log expected wage rate) and inc (income excluding wife’s income), the third model used basis splines on lwg and inc, and the fourth model which is a general additive model using smoothing functions on lwg, and inc. The second model the step function dropped k618 (number of children 6 to 18 years old), wc (wife’s college attendance),hc (husband’s college attendance). After running the third model through the step function it dropped k618 and wc. We then compared the AIC’s and we see the best fit is the polynomial transformations on the lwg and inc. Since the interpretation are difficult due the transformation in the x space we compared the prediction error rate for all 4 models as well.

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in bs(inc, degree = 3L, knots = c(25, 50, 75), Boundary.knots =  
## c(1.5, : some 'x' values beyond boundary knots may cause ill-conditioned  
## bases

## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts  
## argument ignored

## [,1] [,2]   
## ero1 "Model 1" "0.345132743362832"  
## ero2 "Model 2" "0.247787610619469"  
## ero3 "Model 3" "0.305309734513274"  
## ero4 "Model 4" "0.292035398230089"

When comparing the error rates of all four models we see that our second model once again has the lowest error rate out of the four models we fit.

summary(m2)

##   
## Call:  
## glm(formula = lfp ~ k5 + age + poly(lwg, 2, raw = FALSE) + poly(inc,   
## 2, raw = FALSE), family = "binomial", data = Mroz)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.31673 -0.88680 0.06396 0.86360 2.18761   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.46942 0.61873 7.223 5.07e-13 \*\*\*  
## k5 -1.52792 0.22272 -6.860 6.88e-12 \*\*\*  
## age -0.06841 0.01282 -5.336 9.48e-08 \*\*\*  
## poly(lwg, 2, raw = FALSE)1 20.53379 4.27981 4.798 1.60e-06 \*\*\*  
## poly(lwg, 2, raw = FALSE)2 106.23218 12.74395 8.336 < 2e-16 \*\*\*  
## poly(inc, 2, raw = FALSE)1 -9.80916 2.53491 -3.870 0.000109 \*\*\*  
## poly(inc, 2, raw = FALSE)2 3.93619 2.42961 1.620 0.105213   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1029.75 on 752 degrees of freedom  
## Residual deviance: 751.73 on 746 degrees of freedom  
## AIC: 765.73  
##   
## Number of Fisher Scoring iterations: 7

Model 2 gives us the best prediction compared to the 3 other models we tested. We can see that as children over 5 increase we see the log-odds of workforce participation decline by -1.5 and as we see age increase we see the log-odds of workforce participation decline -.06. The conculsion we can draw from our model is that as the number of small childern at home increases, age, and income excluding the husbands salary increase in the first polynmial can all cause declines in the likelihood of labor participation. While my the 1st and 2nd polynimal of log wage indicates that as the expected wage rate increased the likelihood of labor participation increases as well.

### Problem 5 (20 points)

In the mlogit package there is a data set called Fishing. This data depicts a customer choice model of different recreational fishing choices. This is based on price, how many fish are caught, and income. Create a model based on price, catch, and inocme, that tells me about a customers choice of mode. Describe how you arrived at your model and any insights it provides.

model1 <- multinom(mode~., data=Fishing)

## # weights: 44 (30 variable)  
## initial value 1638.599935   
## iter 10 value 1306.158295  
## iter 20 value 1150.757958  
## iter 30 value 1081.847707  
## iter 40 value 1072.832724  
## iter 50 value 1064.623922  
## final value 1064.196788   
## converged

summary(model1)

## Call:  
## multinom(formula = mode ~ ., data = Fishing)  
##   
## Coefficients:  
## (Intercept) price.beach price.pier price.boat price.charter  
## pier -6.252015 -0.002469672 -0.002469672 -0.2934729 0.2936538  
## boat -13.127751 0.015698054 0.015698054 -0.5650396 0.5422299  
## charter -15.581916 0.013459039 0.013459039 -0.6534132 0.6358946  
## catch.beach catch.pier catch.boat catch.charter income  
## pier 34.71251 -48.15681 -1.4650481 0.05204003 -8.987710e-05  
## boat 57.12954 -84.59076 -0.4383507 0.30151639 1.064670e-05  
## charter 58.83028 -84.82774 0.5562081 0.09880167 -6.056782e-05  
##   
## Std. Errors:  
## (Intercept) price.beach price.pier price.boat price.charter  
## pier 0.000504553 0.001672146 0.001672146 0.008273766 0.007445115  
## boat 0.002612635 0.001693029 0.001693029 0.009458203 0.008891930  
## charter 0.002665084 0.001674973 0.001674973 0.009069802 0.008525760  
## catch.beach catch.pier catch.boat catch.charter income  
## pier 0.0001769799 0.0001480912 0.0002019135 0.0007322931 5.019554e-05  
## boat 0.0024047937 0.0016138224 0.0116042051 0.0395411226 5.711061e-05  
## charter 0.0024759877 0.0016492126 0.0118472123 0.0404433209 5.666854e-05  
##   
## Residual Deviance: 2128.394   
## AIC: 2182.394

set.seed(4567)  
trainingIndex <- sample(1:nrow(Fishing), 0.7\*nrow(Fishing))  
trainingData <- Fishing[trainingIndex,]  
testingData <- Fishing[-trainingIndex,]  
  
multinomModel1 <- multinom(mode ~ ., data=trainingData)

## # weights: 44 (30 variable)  
## initial value 1146.465437   
## iter 10 value 943.005349  
## iter 20 value 808.166602  
## iter 30 value 765.286907  
## iter 40 value 760.603177  
## iter 50 value 754.688126  
## final value 753.245289   
## converged

summary (multinomModel1)

## Call:  
## multinom(formula = mode ~ ., data = trainingData)  
##   
## Coefficients:  
## (Intercept) price.beach price.pier price.boat price.charter  
## pier -6.555457 -0.004186186 -0.004186186 -0.3202127 0.3212117  
## boat -12.662975 0.013662762 0.013662762 -0.5391135 0.5190360  
## charter -14.771718 0.011215678 0.011215678 -0.6257665 0.6102915  
## catch.beach catch.pier catch.boat catch.charter income  
## pier 33.78297 -48.35832 -1.712423 0.1163870 -1.173264e-04  
## boat 54.54172 -80.83956 -1.056356 0.4333003 3.304643e-05  
## charter 55.80931 -80.72966 -1.012676 0.4801099 -5.679491e-05  
##   
## Std. Errors:  
## (Intercept) price.beach price.pier price.boat price.charter  
## pier 0.000608011 0.001899614 0.001899614 0.009937239 0.009105234  
## boat 0.003125636 0.001757202 0.001757202 0.011370163 0.010937438  
## charter 0.003190613 0.001728664 0.001728664 0.010733773 0.010296138  
## catch.beach catch.pier catch.boat catch.charter income  
## pier 0.0002232392 0.0001855608 0.0002877286 0.001043672 6.196858e-05  
## boat 0.0027423751 0.0018849099 0.0138604571 0.047083606 6.935291e-05  
## charter 0.0028501272 0.0019443219 0.0142217460 0.048428313 6.847683e-05  
##   
## Residual Deviance: 1506.491   
## AIC: 1560.491

### Interpretation of Coefficients:

Baseline of the multinomial fishing model is beach and therefore estimated a model for pier relative to beach, a model for boat relative to beach and a model for charter relative to beach.

With increase in price for beach mode, individuals are likely to choose boat mode with little change in charter mode. With increase in catch rate for beach mode, individuals are likely to choose charter mode and less likely to choose beach and pier mode.

With increase in price for pier mode, individuals are likely to choose boat mode with little change in charter mode. With increase in cach rate for pier mode, individuals are likely to choose beach mode.

With increase in price for boat mode, individuals are likely to choose beach mode. With increase in catch rate for boat, individuals are likely to choose charter mode.

With increase in price for charter mode, individuals are likely to chooose charter mode. With increase in catch price for charter, individuals are likely to choose boat mode with too little change in charter mode.

With increase in income, there is not much siginificant change in selecting the fishing mode. Minimal change is seen in charter and boat mode.

Looking at coefficients of summary, as the tuning parameters (predictors) changes by one unit, the log odds will decrease by that unit. For example, if price of pier mode changes by -0.0024 and all others are set to zero than response variable will change by -3.2446 unit ( -3.247 - (-0.0024)) against beach mode. If predictors are set to zero, log odds of pier mode is -3.2473 against beach mode, log odds of boat mode is -10.288 against beach mode and log odds of charter mode is -12.8385 against beach mode. In general, if a unit change in the predictor variable, the intercept to the baseline (beach mode) is expected to change by its respective parameter estimate, which is in log-odds units given the variables in the model are held constant.

Predicting the probablities of variables at different values,all the variables are making difference on the response. Moreover, effects plot shows similar visibility for choosing the fishing mode. The model accuracy with test data is 73%.

### Conclusions Drawn:

After looking and interpreting the coefficients and running through prediction cases, we decided that we cannot draw any relationship conclusions from the dataset. There doesn’t seem to be any correlation between the fishing mode choice based on price, catch, and income. Looking at catch rate it seems individuals catch a lot more on average on the charter compared to other fishing modes. This can explain why even though the charter.price goes up, individuals will still pick the charter over the beach. Even if the catch.beach goes up individuals are still choosing the charter over the beach because 1 unit increase of catch.beach isn’t proportional to one unit of catch.charter increase.

We believe the main culprit for not being able to draw any conclusions on an individual’s choice of fishing mode is the dataset. The data is said to be from individuals across the US. However, it doesn’t give us the specific location of the indivdual (region, state, beach) or time of the year. Both location and time of year can have big impacts on availabilty of boats and charters. We also do not get information from the sample of the size of the fishing parties on the boat or charters. The interpretations could change depending on if the individuals are out fishing alone or with large groups of people. These are all factors that would have a greater impact than price, catch, and income. Taking into account the data summary, model coefficents, prediction models/cases and the data issues we see, we believe an individul’s choice was more based on their situation than on price, catch, and income. There is no hard conclusion that can be drawn for someone choosing a specific fishing mode over another based solely on price, catch, and income from the information in this dataset.