BAS 320 - Assignment 9 - Logistic Regression

Jack Guptill

## Analysis of the Probability of Travelers Buying Travelers Insurance based on various Predictor Variables

I’m using a logistic regression model to predict the probability that an individual buys travelers insurance (Yes level) from what the travelers age is, whether they graduated whether they are employed by the government or if they work in the private sector/self-employed (combined level), what their annual income is, whether or not they have a chronic disease, if they are a frequent flyer, if they have ever traveled abroad, and if they have 2-4, 5, or 6-9 family members.

I’m making this model because I am genuinely interested in the underwriting and risk analysis of insurance companies and I thought that this would be a good introduction to the data that insurance companies use to analyze policy holders. I am curious to what sorts of people buy travelers insurance in general.

The data I’m using comes from Kaggle (<https://www.kaggle.com/datasets/tejashvi14/travel-insurance-prediction-data>) and contains a total of 1987 rows and 8 total predictors and 1 target variable.

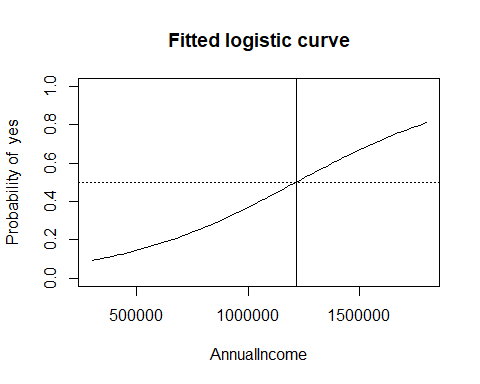
## Task 1 - Investigation of the relationship between the probability of a Traveler buying Travelers Insurance and and their Annual Income

Those that have a higher annual income have a larger probability of buying travelers income. Conversely, those that make less per year have a lower probability of buying travelers insurance. The 50/50 break point for this data set is with those that make $1,215,900. In other words, for those that make around $1.2 million per year have a 50/50 chance of buying travelers insurance.

M.task1 <- glm(TravelInsurance~AnnualIncome, data=DATA, family=binomial) #Fit a simple logistic regression predicting Y from X (numeric)  
summary(M.task1)

##   
## Call:  
## glm(formula = TravelInsurance ~ AnnualIncome, family = binomial,   
## data = DATA)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.013e+00 1.608e-01 -18.73 <2e-16 \*\*\*  
## AnnualIncome 2.478e-06 1.496e-07 16.56 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2590.5 on 1986 degrees of freedom  
## Residual deviance: 2259.5 on 1985 degrees of freedom  
## AIC: 2263.5  
##   
## Number of Fisher Scoring iterations: 3

visualize\_model(M.task1);abline(v=3.013/.000002478)



#You'll also need one calculation here that gets the value of X that has a 50% chance of having the level of interest  
3.013/.000002478 #1215900

## [1] 1215900

## Task 2 - Investigation of the Relationship Between the Probability of Travelers Buying Travelers Insurance by whether they are Frequent Flyers and what their Annual Income is

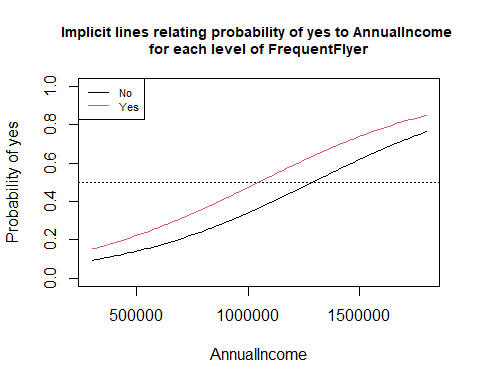
From looking at the output we can tell that travelers that are frequent flyers have a higher probability of flying that those that are not frequent flyers.

Travelers that are frequent flyers have a stronger relationship between the probability of getting insurance and annual income. Over the range of annual incomes that we see in the data, the probability of those that are not frequent flyers that make less than approximately 800K in annual income is higher than those that are frequent flyers. This flips however, with those that make ~$800K. In other words, those that make above ~800K and are frequent fliers are much more likely to buy travelers insurance.

M.task2.noint <- glm(TravelInsurance~FrequentFlyer+AnnualIncome, data=DATA, family=binomial) #Fit a regression predicting Y from X1 (numeric) and X2 (categorical) that EXCLUDES the interaction  
summary(M.task2.noint )

##   
## Call:  
## glm(formula = TravelInsurance ~ FrequentFlyer + AnnualIncome,   
## family = binomial, data = DATA)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.947e+00 1.617e-01 -18.219 < 2e-16 \*\*\*  
## FrequentFlyerYes 5.532e-01 1.258e-01 4.397 1.1e-05 \*\*\*  
## AnnualIncome 2.288e-06 1.560e-07 14.668 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2590.5 on 1986 degrees of freedom  
## Residual deviance: 2240.4 on 1984 degrees of freedom  
## AIC: 2246.4  
##   
## Number of Fisher Scoring iterations: 3

visualize\_model(M.task2.noint )

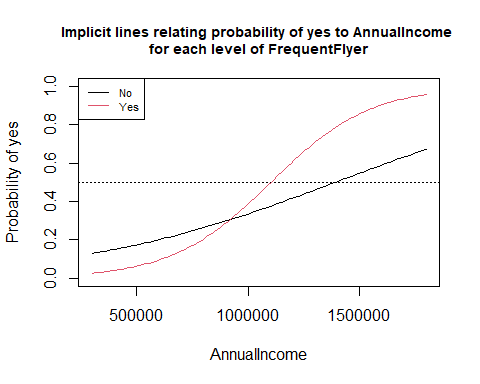


##   
## Effect test for FrequentFlyer has p-value 1.184e-05

M.task2.int <- glm(TravelInsurance~FrequentFlyer\*AnnualIncome, data=DATA, family=binomial) #Fit a regression predicting Y from X1 (numeric) and X2 (categorical) that INCLUDES the interaction  
summary(M.task2.int )

##   
## Call:  
## glm(formula = TravelInsurance ~ FrequentFlyer \* AnnualIncome,   
## family = binomial, data = DATA)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.428e+00 1.710e-01 -14.199 < 2e-16 \*\*\*  
## FrequentFlyerYes -2.516e+00 5.248e-01 -4.795 1.62e-06 \*\*\*  
## AnnualIncome 1.745e-06 1.706e-07 10.231 < 2e-16 \*\*\*  
## FrequentFlyerYes:AnnualIncome 2.738e-06 4.469e-07 6.127 8.94e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2590.5 on 1986 degrees of freedom  
## Residual deviance: 2195.5 on 1983 degrees of freedom  
## AIC: 2203.5  
##   
## Number of Fisher Scoring iterations: 4

visualize\_model(M.task2.int )



##   
## Effect test for interaction with FrequentFlyer has p-value 2.09e-11

TO.PREDICT <- data.frame(  
 AnnualIncome = c(600000,1250000),  
 FrequentFlyer=c('No','Yes')  
) #Construct the dataframe where you're making predictions; 2 rows  
predict(M.task2.int,newdata=TO.PREDICT,type="response")

## 1 2   
## 0.2009234 0.6592499

## Task 3 - Investigation of the relationship between Travelers buying Travelers Insurance and all Predictor Variables in the Dataset (Full Model)

Two travelers that are otherwise identical in their employment type, whether they graduated or not, their age, the amount of family members they have, if they have chronic diseases, if they are a frequent flyer or not, if they have ever traveled abroad that differ in annual income, the traveler that has a higher annual income has a higher probability of buying travelers insurance. With this variable being statistically significant we now know that a travelers annual income brings additional information to the model in understanding the probability of them buying travelers insurance above and beyond what the information that the other variables bring into the model.

Among other wise identical travelers, those that either work in the private sector or are self employed have a higher probability of buying travelers insurance then does those that are working in the government sector.

Among otherwise identical travelers, the traveler that has a family size of 6 or greater (5 exclusive and above in the output) has the highest probability of getting travelers insurance. Those that have a family size of 5 have a lower probability of getting travelers insurance than those that have a family size of 2-4. For this analysis we are using the drop 1 method in seeing if the variable, family size, is statistically significant if it is taken out of the model. Given that family size is statistically signifant when taken out of the model, we know that adding it into the model brings in additional new information that other variables cannot bring in by themselves.

DATA$FamilyMembers <- NULL  
DATA$AgeCat = NULL

M.full <- glm(TravelInsurance~.,data=DATA, family=binomial) #Fit a regression predicting Y from everything  
summary(M.full)

##   
## Call:  
## glm(formula = TravelInsurance ~ ., family = binomial, data = DATA)  
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -4.857e+00 6.181e-01 -7.857  
## Age 7.316e-02 1.861e-02 3.931  
## Employment.TypePrivate Sector/Self Employed 1.110e-01 1.334e-01 0.832  
## GraduateOrNotYes -1.725e-01 1.573e-01 -1.097  
## AnnualIncome 1.570e-06 1.780e-07 8.821  
## ChronicDiseases1 9.067e-02 1.217e-01 0.745  
## FrequentFlyerYes 4.803e-01 1.377e-01 3.488  
## EverTravelledAbroadYes 1.727e+00 1.540e-01 11.214  
## FamilyMemberCat(4,5] -1.922e-01 1.466e-01 -1.311  
## FamilyMemberCat(5,9] 6.411e-01 1.244e-01 5.155  
## Pr(>|z|)   
## (Intercept) 3.92e-15 \*\*\*  
## Age 8.45e-05 \*\*\*  
## Employment.TypePrivate Sector/Self Employed 0.405305   
## GraduateOrNotYes 0.272724   
## AnnualIncome < 2e-16 \*\*\*  
## ChronicDiseases1 0.456161   
## FrequentFlyerYes 0.000487 \*\*\*  
## EverTravelledAbroadYes < 2e-16 \*\*\*  
## FamilyMemberCat(4,5] 0.189987   
## FamilyMemberCat(5,9] 2.54e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2590.5 on 1986 degrees of freedom  
## Residual deviance: 2051.7 on 1977 degrees of freedom  
## AIC: 2071.7  
##   
## Number of Fisher Scoring iterations: 4

drop1(M.full, test="Chisq") #reference level is [2,4] or 2 to 4 family members

## Single term deletions  
##   
## Model:  
## TravelInsurance ~ Age + Employment.Type + GraduateOrNot + AnnualIncome +   
## ChronicDiseases + FrequentFlyer + EverTravelledAbroad + FamilyMemberCat  
## Df Deviance AIC LRT Pr(>Chi)   
## <none> 2051.7 2071.7   
## Age 1 2067.2 2085.2 15.505 8.230e-05 \*\*\*  
## Employment.Type 1 2052.4 2070.4 0.695 0.4045524   
## GraduateOrNot 1 2052.9 2070.9 1.193 0.2747673   
## AnnualIncome 1 2132.3 2150.3 80.629 < 2.2e-16 \*\*\*  
## ChronicDiseases 1 2052.2 2070.2 0.554 0.4568320   
## FrequentFlyer 1 2063.7 2081.7 12.008 0.0005297 \*\*\*  
## EverTravelledAbroad 1 2189.1 2207.1 137.461 < 2.2e-16 \*\*\*  
## FamilyMemberCat 2 2089.1 2105.1 37.460 7.341e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Task 4 - Assessing the full model

The naive model’s accuracy is 64% by using the majority level. The first model that doesn’t include interactions has an accuracy of 77.52% and the model that does include interactions has an accuracy of 81.25%. This second model with the interactions has the highest overall accuracy.

In terms of our full model that doesn’t include all interactions between all of the variables, it is not a reasonable reflection of reality as the model does not pass both goodness of fit tests. This model might still be good to be making predictions, however the sigmoidal curve doesn’t seem to be a good representation of the probability of determining if someone is going to buy travelers insurance.

With regards to the second model that does include all two way interactions, it does seem to somewhat okay in terms of being a reasonable reflection of reality as it does barely pass the first test but it fails the second. Again this model still might be good to make predictions and still might be somewhat useful.

M.full2 <- glm(TravelInsurance~.^2, data=DATA, family=binomial) #Fit a regression predicting Y from all predictors and all interactions  
  
confusion\_matrix(M.full) #77.52% accuracy

## Predicted no Predicted yes Total  
## Actual no 1175 102 1277  
## Actual yes 350 360 710  
## Total 1525 462 1987

confusion\_matrix(M.full2) #81.25% accuracy

## Predicted no Predicted yes Total  
## Actual no 1192 85 1277  
## Actual yes 282 428 710  
## Total 1474 513 1987

table(DATA$TravelInsurance) #You need a frequency table of the column containing Y and a calculation of the naive model's accuracy

##   
## no yes   
## 1277 710

1277/(1277+710) #64% naive accuracy

## [1] 0.6426774

check\_regression(M.full)

## Method 1 (comparing each observation with simulated results given model is correct; not very sensitive)  
## p-value of goodness of fit test is approximately 0.044  
##   
## Method 2 (Hosmer-Lemeshow test with 10 categories; overly sensitive for large sample sizes)   
## p-value of goodness of fit test is approximately 0

check\_regression(M.full2)

## Method 1 (comparing each observation with simulated results given model is correct; not very sensitive)  
## p-value of goodness of fit test is approximately 0.058  
##   
## Method 2 (Hosmer-Lemeshow test with 10 categories; overly sensitive for large sample sizes)   
## p-value of goodness of fit test is approximately 0.01