Cryotherapy as a Treatment Method

Math 4322

Group 13: Jacob Kilgore, Mihir Pingili

2 Dec. 2022

Introduction

In the medical field there are many different forms of treatment available to those who have growths on the surface of the skin, one such treatment being cryotherapy. It is a simple non-evasive procedure which uses liquid nitrogen to freeze and destroy any precancerous skin conditions such as warts, skin tags and moles at the cellular level. This treatment method is affordable with a low level of discomfort and risk of infection and is gaining traction to become of the primary methods for removal of benign surface level skin growths.

We are using the UCI Cryotherapy Dataset for this study. The data has observations of the 90 different patients who underwent a form of cryotherapy treatment using liquid nitrogen for warts. This information can be used to help show if cryotherapy is a good treatment method given it is a relatively newer medical practice that has only recently seen more common clinical use. Additionally, being able to predict whether cryotherapy will be successful or not based on the details of a patient will help calculate the risk involved, which most medical procedures, cryotherapy included, have to some extent. As different applications and more research is done towards cryotherapy and medicine in general, being able to analyze the efficacy of the treatments will be crucial to safely advancing practices such as this.

The dataset contains 90 observations with 7 variables each and missing no data values:

- Result of treatment - successful (0) or unsuccessful (1): categorical
- Sex - male (1) or female (2): categorical
- Age - in years: continuous
- Time - length of time before the treatment in months: continuous
- Number of warts - continuous
- Type - Type of wart being treated – Common (1) or Plantar (2) or Both (3): categorical
- Area - Surface area of the warts in mm^2: continuous

We want to prioritize prediction as it a classification problem with our response variable being the result of treatment and the other variables serving as predictor. For our models we chose to use logistic regression and random forests for reasons discussed later in the study.

We will examine the data set to (1) predict whether cryotherapy treatment will be successful or not; (2) which variable predictors demonstrate a statistically significant relationship to the result of the treatment.

Logistic Regression (Jacob Kilgore)

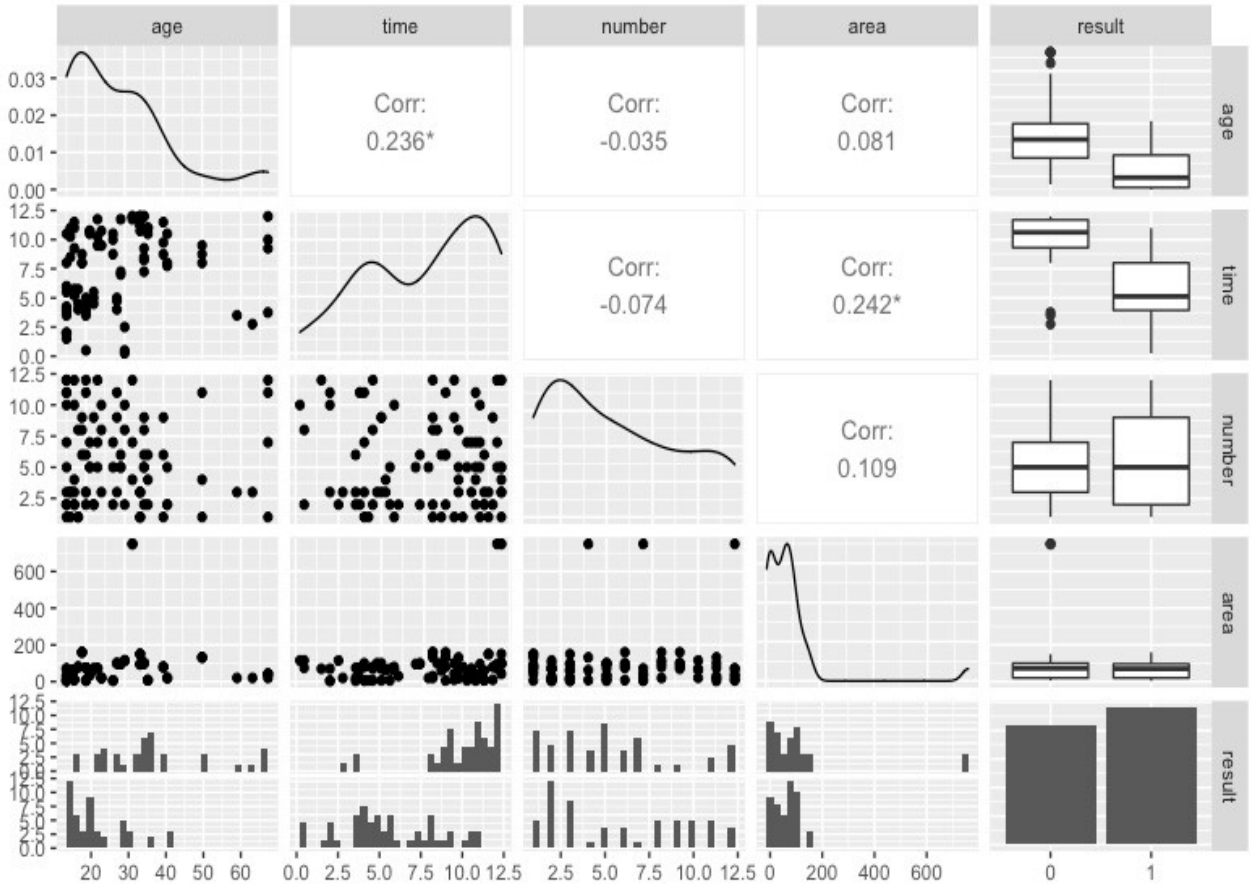
We are using logistic regression to help with this as the dataset has a somewhat small number of observations and we want the observations to be easy to understand so logistic regression seems like a perfect fit that can also serve as a good baseline to start understanding the dataset and compare to the other model we're using.

The model for our formula is

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}$$

with the β's being the coefficients and the X's being some or all of our predictors depending on which we include.

```
Cryotherapy$result = factor(Cryotherapy$result)
ggpairs(Cryotherapy[,c(2:4,6,7)])
```



Initial model:

```
cryo.glm = glm(result ~ age + time + sex + number + type + area, data = Cryot
herapy, family = "binomial")
summary(cryo.glm)
```

```
##
## Call:
## glm(formula = result ~ age + time + sex + number + type + area,
##     family = "binomial", data = Cryotherapy)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.14221  -0.07647   0.00427   0.08912   2.33962
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 17.894370   5.770302   3.101  0.00193 **
## age         -0.156638   0.065954  -2.375  0.01755 *
## time        -1.472689   0.456677  -3.225  0.00126 **
## sex2        -1.992336   1.434853  -1.389  0.16498
## number       0.039908   0.189449   0.211  0.83316
## type2        3.033105   1.677794   1.808  0.07064 .
## type3       -6.959380   3.387303  -2.055  0.03992 *
## area         0.010024   0.007967   1.258  0.20833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 124.366  on 89  degrees of freedom
## Residual deviance:  24.328  on 82  degrees of freedom
## AIC: 40.328
##
## Number of Fisher Scoring iterations: 8
```

Final model:

```
cryo.glm = glm(result ~ age + time + type, data = Cryotherapy, family = "bino
mial")
summary(cryo.glm)
```

```
#
## Call:
## glm(formula = result ~ age + time + type, family = "binomial",
##     data = Cryotherapy)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.79299  -0.05346   0.00436   0.07536   1.90310
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 16.36796    4.65840   3.514 0.000442 ***
## age         -0.13816    0.05709  -2.420 0.015525 *
## time        -1.36599    0.41088  -3.325 0.000886 ***
## type2        2.27367    1.29166   1.760 0.078361 .
## type3       -6.76688    3.40370  -1.988 0.046801 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 124.366  on 89  degrees of freedom
## Residual deviance:  28.636  on 85  degrees of freedom
## AIC: 38.636
##
## Number of Fisher Scoring iterations: 8
```

I excluded the Number of Warts, Sex, and Area because after fitting them to the model, they had a large p-value and it was clear that they didn't improve the accuracy of the model and in Number of Warts & Area's case, the pair plots didn't indicate much of a correlation with the target variable for either. Moreover, Number of Warts wouldn't impact the effect much given how the cryotherapy treatment is applied.

```
err = rep(0,10)
for(i in 1:10)
{
  set.seed(i)
  sample = sample.int(n = nrow(Cryotherapy), size = floor(.8*nrow(Cryotherapy
)), replace=F)
  training = Cryotherapy[sample,]
  testing = Cryotherapy[-sample,]
  cryo.glm = glm(result ~ age + time + type, data = training, family = "binom
ial")
  cryopred = predict.glm(cryo.glm, newdata=testing, type="response")
  yhat = cryopred > 0.5
  conf.mat = table(yhat,testing$result)
  err[i] = (conf.mat[1,2]+conf.mat[2,1])/18
}
mean(err)
```

```
## [1] 0.1055556
```

The average test prediction error for the 10 iterations of our model for logistic regression was 10.5% and given the repeated cross validation we use, it's safe to say that the baseline model we've made does a good job at predicting the result of the treatment while also giving some easily understandable insight as to the underlying patterns in the data.

Random Forests (Mihir Pingili)

The most used and best supervised learning models are considered to be tree-based methods. They are adaptable and can map non-linear relationships well. Methods such as decision trees are simple and easy to understand but have lower prediction accuracy when compared to other models and suffer from high variance where different subsets of the same data could produce drastically different results. To fix this we use the bootstrapping method on decision trees where data sets are created to obtain a single general prediction which improves the prediction accuracy. To do this we can use Bagging (bootstrap aggregation) but it can introduce a loss of interpretability of a model and can experience bias. Correlation of the bagged trees can lead to the same variable largely influencing each bagged tree. To improve on this, we use Random Forests which permits a reduction in variance and consequently a lower test error rate. Random Forests displays the same advantages of single decision tree models and bagging with its capability to work on continuous, categorical, and even non-linear relationships. Due to there being a large number of trees it increases the complexity and hence requires much more computational power and resources and also have a longer training period than decision trees.

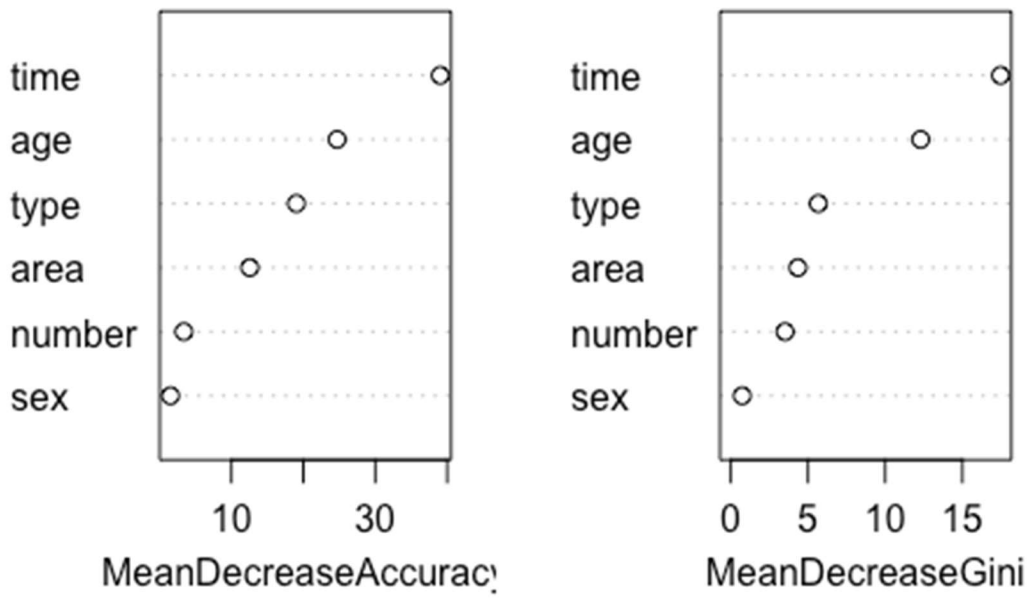
The model formula is $Result \sim Sex + Age + Time + Number + Type + Area$

To start building the model, we create an initial random forest model (`prelim.rf`) using all six predictors.

```
prelim.rf = randomForest(result ~ sex + age + time + number + type + area,
                         data = Cryotherapy, mtry = sqrt(6), importance = TRUE)
prelim.rf

##
## Call:
##  randomForest(formula = result ~ sex + age + time + number + type + area,
data = Cryotherapy, mtry = sqrt(6), importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 6.67%
## Confusion matrix:
##    0  1 class.error
## 0 41  1  0.02380952
## 1  5 43  0.10416667

varImpPlot(prelim.rf, sort = TRUE, main=NA)
```



The initial model lists the call used to build the classifier, the number of trees (500), the variables at each split (2). The OOB (out-of-bag) estimate of error rate is a measure to distinguish between different random forest classifiers, to produce a lower value for error rate we could vary the number of trees or number of variables. It is calculated using the confusion matrix by counting how many points were misclassified (5+1=6) and dividing this by the total number of observations (6/90 ≈ 6.67%).

The importance classifier allows us to plot and visualize the decision making process using `varImpPlot()`and assess each feature based on the two criteria of *MeanDecreaseAccuracy* and *MeanDecreaseGini. MeanDecreaseAccuracy* gives an estimate of the loss in prediction performance when each variable is excluded. *MeanDecreaseGini* is the total decrease in node impurity from splitting the variable, averaged over all trees. The higher the value of either variable, the higher the importance of the variable in the model. From the *MeanDecreaseAccuracy* plot `time`, `age`, `type`, and `area` seem to have a significant relationship to the result of the treatment while `number`, and `sex` have very little effect on prediction performance and insignificant enough to be omitted. Similarly, from the *MeanDecreaseGini* plot `time` and `age` are the top two predictors while `type`, `area`, `number`, and `sex` are at the bottom having high node impurity and hence low significance to the predictive power of the model.

From looking at both plots it is clear that `number`, and `sex` are the least significant to the random forest model as they were the lowest performing predictors in both metrics and hence were ranked lower. While `time` and `age` were the clear top two predictors in both plots, `type` and `area` were not as significant and in the *MeanDecreaseGini* plot they did not differ much from `time` and `age`. As `type` and `area` appeared relevant in the *MeanDecreaseAccuracy* plot they will be included in the model. The revised model (`result.rf`) will have a model formula of: *Result ~ Age + Time + Type + Area*. The mtry was not changed as changing it would not have helped as the number of predictors is already low.

```
result.rf = randomForest(result ~ age + time + type + area,
                         data = Cryotherapy, mtry = sqrt(6), importance = TRUE)
result.rf

##
## Call:
##  randomForest(formula = result ~ age + time + type + area, data = Cryother
apy,      mtry = sqrt(6), importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 4.44%
## Confusion matrix:
##    0  1 class.error
## 0 42  0  0.00000000
## 1  4 44  0.08333333
```

In this model we obtained an error rate of 4.44% which shows an improvement from the initial model (`prelim.rf`) which had an error rate of 6.67%. Seeing as the error rate was reduced, we can proceed to train the model using a randomized 80:20 split to train using 80% of the data to train the model and the other 20% to test the model.

```
error.rate = rep(0,10)
for(i in 1:10)
  {
    set.seed(i)
    index = sample.int(n = nrow(Cryotherapy), size = floor(.8*nrow(Cryotherap
y)), replace=F)
    train = Cryotherapy[index,]
    test = Cryotherapy[-index,]
    result.rf = randomForest(result ~ age + time + type + area,
                             data = train, mtry = sqrt(6), importance = TRUE)
```

```
    yhat.rf = predict(result.rf, newdata=test)
    con.matrix = table(yhat.rf,test$result)
    error.rate[i] = (con.matrix[1,2]+con.matrix[2,1])/sum(con.matrix)
}
error.rate
```

```
##  [1] 0.22222222 0.11111111 0.05555556 0.11111111 0.00000000 0.05555556
##  [7] 0.00000000 0.16666667 0.11111111 0.11111111
```

```
mean(error.rate)
```

```
## [1] 0.09444444
```

We obtained an average error rate of 9.44% across all 10 iterations of training and testing of the random forest model, with error rates ranging between 0.00 and 0.222.

Conclusion

Both of our models served well in predicting whether or not the treatment could be effective. The predictors we used in each of the models were similar, aside from the inclusion of area in the random forest model. The random forest model performed slightly better than the logistic regression model with 9.44% and 10.5% respectively. This falls in line with expectations due to random forests being more complex with the use of bootstrapping decision trees and the ability to reduce variance, minimize overfitting and are typically more accurate when compared to a simple logistic regression model.

For the purpose of understanding the models, logistic regression is more easily interpretable and allowed us to determine the influence each predictor variable had on the response variable. We found that time, age, and type of wart were strong predictors while number of warts, sex, and area were insignificant in predicting the outcome of the treatment. This is accordant to the random forest models findings which also used the area predictor variable in the final model as removing the variable during trial and error caused an increase in the error rate as opposed to when it was included. Time having the highest influence is surprising as all the warts are benign. As for the main point of figuring out whether we can predict the success of the treatment or not and which predictors are the most useful in achieving this goal, we achieved both and either model could be used.

Cryotherapy as a treatment method being relatively new, we were able to create models that predict the outcome of the treatment well with the small dataset we were able to find. As

more data is collected and new variables are introduced as possible predictors it is possible to improve on the models and achieve much lower error rates with the same type of models. To further improve on our study, we could use Neural networks to create a model to predict the outcome of the treatment but due to the lack of number of observations in the dataset the model may not perform better than the two models used in this study.

Bibliography

- F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, and S. Nahavandi, "An expert system for selecting wart treatment method," Computers in Biology and Medicine, vol. 81, pp. 167-175, 2/1/ 2017.

- F. Khozeimeh, F. Jabbari Azad, Y. Mahboubi Oskouei, M. Jafari, S. Tehranian, R. Alizadehsani, et al., "Intralesional immunotherapy compared to cryotherapy in the treatment of warts," International Journal of Dermatology, 2017, DOI: 10.1111/ijd.13535

- *Cryotherapy Wart & Mole removal*. U.S. Dermatology Partners. (2021, June 25). Retrieved December 2, 2022, from https://www.usdermatologypartners.com/services/skin-cancer/cryotherapy/#:~:text=Cryotherapy%2C%20or%20%E2%80%9Ccryosurgery%2C%E2%80%9D,warts%2C%20skin%20tags%20and%20moles.

- *Computers in biology and medicine - isi articles*. (n.d.). Retrieved November 30, 2022, from https://isiarticles.com/bundles/Article/pre/pdf/138740.pdf

- Andrews, M. D. (2004, May 15). *Cryosurgery for common skin conditions*. American Family Physician. Retrieved November 30, 2022, from https://www.aafp.org/pubs/afp/issues/2004/0515/p2365.html