Georgia Tech

WEEK 5 HOMEWORK

INSTRUCTIONS

- Every learner should submit his/her own homework solutions. However, you are allowed to discuss the homework with each other (in fact, I encourage you to form groups and/or use the forums) – but everyone must submit his/her own solution; you may not copy someone else's solution.
- The homework will be peer-graded. In analytics modeling, there are often lots of different approaches that work well, and I want you to see not just your own, but also others.
- The homework grading scale reflects the fact that the primary purpose of homework is learning:

Rating	Meaning	Point value (out of 100)
4	All correct (perhaps except a few details) with a deeper solution than expected	100
3	Most or all correct	90
2	Not correct, but a reasonable attempt	75
1	Not correct, insufficient effort	50
0	Not submitted	0

Question 11.1

Using the crime data set `uscrime.txt` from Questions 8.2, 9.1, and 10.1, build a regression model using:

1. Stepwise regression
2. Lasso
3. Elastic net

For Parts 2 and 3, remember to scale the data first – otherwise, the regression coefficients will be on different scales and the constraint won't have the desired effect.

For Parts 2 and 3, use the `glmnet` function in R.

Notes on R:

- For the elastic net model, what we called λ in the videos, `glmnet` calls "alpha"; you can get a range of results by varying alpha from 1 (lasso) to 0 (ridge regression) [and, of course, other values of alpha in between].
- In a function call like `glmnet(x,y,family="mgaussian",alpha=1)` the predictors `x` need to be in R's matrix format, rather than data frame format. You can convert a data frame to a matrix using `as.matrix` – for example, `x <- as.matrix(data[,1:n-1])`
- Rather than specifying a value of T, `glmnet` returns models for a variety of values of T.

Before fitting the models, we need to preprocess the data appropriately. This includes scaling the continuous predictors to ensure they are on a comparable scale.
Scaling the data except the response variable that is Crime variable and categorical variable “So”

Defining control using 5-fold cross validation
train_control <- trainControl(method = "cv", number = 5)

Stepwise Regression
we employ backward elimination, where we start with all potential predictors and iteratively remove the least significant ones.

```
stepwise_model <- train(Crime ~ ., data = scaledData, method = "lmStepAIC", trControl =
train_control, direction = "both")
summary(stepwise_model$finalModel)
```

The backward stepwise regression selected eight significant variables: M.F, U1, Prob, U2, M, Ed, Ineq, and Po1. The adjusted R-squared value for this model was 0.7444, indicating that approximately 74.44% of the variability in the crime rate can be explained by these predictors.

Lasso Regression

```
XP = data.matrix(scaledData[,-16])
YP = data.matrix(scaledData$Crime)

lasso_model <- cv.glmnet(x = XP, y = YP, alpha = 1, nfolds = 5, type.measure = "mse", family =
"gaussian")
best_lambda <- lasso_model$lambda.min

lasso_coef <- as.matrix(coef(lasso_model, s = best_lambda))
print(lasso_coef)

lasso_vars <- rownames(lasso_coef)[lasso_coef != 0]
lasso_formula <- as.formula(paste("Crime ~", paste(lasso_vars[-1], collapse = " + ")))
mod_lasso <- lm(lasso_formula, data = scaledData)
summary(mod_lasso)
```

The Lasso model selected nine predictors: So, M, Ed, Po1, M.F, NW, U2, Ineq, and Prob. The adjusted R-squared value for the Lasso model was slightly lower at 0.72, suggesting that this model is slightly less explanatory compared to the stepwise regression model.

Elastic Net Regression

```
R2 <- c()
for (i in 0:10) {
  model <- cv.glmnet(x = XP, y = YP, alpha = i / 10, nfolds = 5, type.measure = "mse", family = "gaussian")
  R2 <- cbind(R2, model$glmnet.fit$dev.ratio[which(model$glmnet.fit$lambda == model$lambda.min)])
}).

alpha_best <- (which.max(R2) - 1) / 10
elastic_net_model <- cv.glmnet(x = XP, y = YP, alpha = alpha_best, nfolds = 5, type.measure = "mse",
family = "gaussian")
best_lambda_en <- elastic_net_model$lambda.min
elastic_net_coef <- as.matrix(coef(elastic_net_model, s = best_lambda_en))
print(elastic_net_coef)
elastic_net_vars <- rownames(elastic_net_coef)[elastic_net_coef != 0]
elastic_net_formula <- as.formula(paste("Crime ~", paste(elastic_net_vars[-1], collapse = " + ")))
mod_Elastic_net <- lm(elastic_net_formula, data = scaledData)
summary(mod_Elastic_net)
```

The Elastic Net model, using an optimal alpha value of 1, selected thirteen predictors: So, M, Ed, Po1, Po2, M.F, Pop, NW, U1, U2, Wealth, Ineq, and Prob. However, despite including more variables, the adjusted R-squared value was 0.7219, not significantly higher than the Lasso model but using more variables, which could indicate overfitting.

Model Evaluation
We then cross-validated these models to evaluate their predictive performance.

```
cross_validation_R2 <- function(model_formula, data) {
  SStot <- sum((data$Crime - mean(data$Crime))^2)
  totsse <- 0
  for(i in 1:nrow(data)) {
    model_i <- lm(model_formula, data = data[-i,])
    pred_i <- predict(model_i, newdata = data[i,])
    totsse <- totsse + ((pred_i - data[i,16])^2)
  }
  R2 <- 1 - totsse / SStot
  return(R2)
      }
```

Cross-validation for Stepwise model

```
stepwise_vars <- names(coef(stepwise_model$finalModel))[-1]
stepwise_formula <- as.formula(paste("Crime ~", paste(stepwise_vars, collapse = " + ")))
R2_Stepwise <- cross_validation_R2(stepwise_formula, scaledData)
cat("Cross-validated R-squared for Stepwise Regression:", R2_Stepwise, "\n")
```

```
Cross-validation for Lasso model
R2_Lasso <- cross_validation_R2(lasso_formula, scaledData)
cat("Cross-validated R-squared for Lasso Regression:", R2_Lasso, "\n")

Cross-validation for Elastic Net model
R2_ElasticNet <- cross_validation_R2(elastic_net_formula, scaledData)
cat("Cross-validated R-squared for Elastic Net Regression:", R2_ElasticNet, "\n")
```

The cross-validated R-squared values were 0.666 for the Stepwise model, 0.62 for the Lasso model, and 0.574 for the Elastic Net model. Despite the initial higher adjusted R-squared values, the Elastic Net model did not perform well on cross-validation, suggesting potential overfitting.

Conclusion
Among the three methods, the Stepwise Regression model provided the best balance between model simplicity and explanatory power, with the highest cross-validated R-squared value. The Lasso model, while slightly less explanatory, was effective in variable selection and resulted in a simpler model compared to the Elastic Net. The Elastic Net model included more predictors but did not significantly improve predictive performance, highlighting the importance of model simplicity and the risk of overfitting.

We can conclude that Stepwise Regression offers a robust and interpretable approach for predicting crime rates with the given dataset.

Question 12.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a design of experiments approach would be appropriate.

Optimizing a Child's Playtime with Toys using DOE. Choosing 3 different types of toys: Building blocks, action figures, puzzle, with different environment: Indoors, outdoor, with different time duration: 15 minutes, 30 minutes, 1 hour and playing alone, with a friend, with a parent.
Randomly assign each combination to different play sessions, ensuring you only change one factor level at a time while keeping others constant. Record the child's engagement and enjoyment during each session using a standardized scale. Use statistical software to analyze the data, identifying which factors and interactions significantly affect the child's engagement and enjoyment.
Generate response surface plots to visualize the optimal conditions.

Question 12.2

To determine the value of 10 different yes/no features to the market value of a house (large yard, solar roof, etc.), a real estate agent plans to survey 50 potential buyers, showing a fictitious house with different combinations of features. To reduce the survey size, the agent wants to show just 16 fictitious houses. Use R's `FrF2` function (in the `FrF2` package) to find a fractional factorial design for this

experiment: what set of features should each of the 16 fictitious houses have? Note: the output of `FrF2` is "1" (include) or "-1" (don't include) for each feature.

```
library(FrF2)
design <- FrF2(nruns=16, nfactors=10)
print(design)
```

```
> print(design)
    A  B  C  D  E  F  G  H  J  K
1  -1 -1  1  1  1 -1 -1 -1 -1  1
2  -1 -1  1 -1  1 -1 -1  1  1 -1
3  -1  1  1 -1 -1 -1  1  1 -1  1
4  -1  1  1  1 -1 -1  1 -1  1 -1
5   1  1  1  1  1  1  1  1  1  1
6   1  1  1 -1  1  1  1 -1 -1 -1
7   1  1 -1  1  1 -1 -1  1 -1 -1
8   1 -1 -1 -1 -1 -1  1 -1 -1 -1
9   1 -1  1  1 -1  1 -1  1 -1 -1
10  1  1 -1 -1  1 -1 -1 -1  1  1
11 -1  1 -1 -1 -1  1 -1  1  1 -1
12 -1 -1 -1  1  1  1  1 -1  1 -1
13  1 -1 -1  1 -1 -1  1  1  1  1
14  1 -1  1 -1 -1  1 -1 -1  1  1
15 -1 -1 -1 -1  1  1  1  1 -1  1
16 -1  1 -1  1 -1  1 -1 -1 -1  1
class=design, type= FrF2
```

Using this approach, we reduce the number of surveys from 1024 to 16. Despite the reduction in survey size, it provides sufficient information to understand the impact of each feature on the market value. By using this fractional factorial design, the real estate agent can efficiently gather valuable data to make informed decisions about which features most significantly affect the market value of a house according to potential buyers.

Question 13.1

For each of the following distributions, give an example of data that you would expect to follow this distribution (besides the examples already discussed in class).

a. Binomial
b. Geometric
c. Poisson
d. Exponential
e. Weibull

a. Binomial Distribution
The number of times an elevator stops at a particular floor during a day, given that it makes 100 trips and has a 10% chance of stopping at that floor on any given trip.

b. Geometric Distribution
The number of phone calls you need to make to reach someone at a hospital to get test results, assuming each call has a constant probability of being answered.

c. Poisson Distribution

The number of people arriving at a hospital emergency room per hour during a snowstorm, assuming the arrivals are independent and occur at a constant average rate.

d. Exponential Distribution: The time until the next phone call is received at a call center, assuming calls come in independently at a constant average rate.

e. Weibull Distribution: The time until computer fails, where the likelihood of failure increases with the duration of usage.