DATA 303/473 Assignment 3

Michael Fry 300570669

Q1 (Deviance test, AIC, test MSE)

(a) **(10 marks)** Fit the model and use `anova()` function to do the deviance test to compare the models. Choose the best model.

```
m1 <- gam(medv ~ rm  + lstat, data = Boston)
m2 <- gam(medv ~ rm  + poly(lstat, df=2), data = Boston)
m3 <- gam(medv ~ rm + age + lstat, data = Boston)
m4 <- gam(medv ~ rm + age + poly(lstat, df=2), data = Boston)

pander(anova(m1,m2, test="F"))
```

Table 1: Analysis of Deviance Table

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
503	15439	NA	NA	NA	NA
502	12684	1	2756	109.1	3.099e-23

```
pander(anova(m3,m4, test="F"))
```

Table 2: Analysis of Deviance Table

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
502	15419	NA	NA	NA	NA
501	12231	1	3188	130.6	4.91e-27

```
pander(anova(m2,m4, test="F"))
```

Table 3: Analysis of Deviance Table

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
502	12684	NA	NA	NA	NA
501	12231	1	452.7	18.55	1.997e-05

Model1 vs Model2: p-value = 3.099e-23. The change from lstat to poly(lstat, df=2) improves the model significantly.

Model3 vs Model4: p-value = 4.91e-27. The change from lstat to poly(lstat, df=2) improves the model significantly.

Model2 and Model4 are nested and therefor Anova can be used to compare. Model2 and Model4 outperformed Model1 and Model3 so only Model2 and Model4 will be compared

Model2 vs Model 4: p-value = 1.997e-05. The addition of age as predictor improves the model significantly.

Overall Model4 is the best model from ANOVA analysis.

(b) **(5 marks)** Calculate `AIC` for each model fitted in (a). Choose the best model using the value of `AIC`.

```
pander(AIC(m1,m2,m3,m4),digits = 3)
```

	df	AIC
m1	4	3174
m2	5	3076
m3	5	3175
m4	6	3060

The AIC values indicate Model4 is the best model as it has the smallest AIC. AIC is >10 smaller than the next model, m2, the more complex model is preferred, model4.

(c) **(10 marks)** Split the data set (100%) into a training set (80%) and a test set (20%). Then fit model1–model4 on the training set, and calculate the test MSE for each model. Choose the best model.

```
train_index <- sample(length(Boston$medv),length(Boston$medv)*0.8)
train <- Boston[train_index,]
test <- Boston[-train_index,]

m1.fit <- gam(medv ~ rm  + lstat, data = train)
m2.fit <- gam(medv ~ rm  + poly(lstat, df=2), data = train)
m3.fit <- gam(medv ~ rm + age + lstat, data = train)
m4.fit <- gam(medv ~ rm + age + poly(lstat, df=2), data = train)

m1.mse <- mean((test$medv - predict(m1.fit, test))^2)
m2.mse <- mean((test$medv - predict(m2.fit, test))^2)
m3.mse <- mean((test$medv - predict(m3.fit, test))^2)
m4.mse <- mean((test$medv - predict(m4.fit, test))^2)

Model_names <- c('Model1', 'Model2', 'Model3','Model4')
Mean_Squared_Error <- c(m1.mse,m2.mse,m3.mse,m4.mse)

pander(data.frame(Model_names,Mean_Squared_Error))
```

Model_names	Mean_Squared_Error
Model1	26.02
Model2	24.2
Model3	26.29
Model4	24.59

The Model2 is the best models since it has the smallest test MSE.

(d) **(10 marks)** By combining the result from (a), (b) and (c), decide the best model. Refit the chosen model using all of the `Boston` data set. Make a prediction of `medv` for a suburb with values `rm=10`, `age=50` and `lstat=10`. Interpret the predicted value.

Both Anova and AIC testing showed Model4 to be the best model. Calculating Test MSE however showed Model1 as the best.

Given that Anova and AIC showed that Model4 was preferred, I will continue to use Model4 as the best

model.

Also, out of the two models Model2 and Model4, only Model4 includes the variable age, so to conduct the prediction asked above, only model4 could be fitted.

```
best_model <- m4
new_data <- data.frame(rm = 10, age = 50, lstat = 10)
prediction <- predict(best_model, newdata = new_data)
pander(prediction)
```

1
36.7

The predicted value of 36.7 is interpreted as;

The predicted median value of owner-occupied homes in a Boston suburb where the average number of rooms per dwelling is 10, the proportion of owner-occupied units built prior to 1940 is 50 and the percent of households with low socioeconomic status is 10, is $36700 (36.7*1000)

Q2 (LASSO, best subset selection)

We continue to work on `Boston` data set. The aim in Q2 is to predict `medv` (median house value) using all predictors in `Boston` data set. In the following questions, we apply LASSO and the best subset selection methods.

(a) **(10 marks)** (LASSO) Fit a lasso model on the training set, with λ chosen by cross-validation with the `1 se rule` . Report the test error obtained, along with the values of non-zero coefficient estimates. We use the training set and the test set created in Q1 (c).

```
# Convert the predictors and response variables to matrix format
x <- model.matrix(medv ~., train)[,-1] # remove intercept
y <- train$medv

lasso.model = glmnet(x,y,alpha=1)

cv.out <- cv.glmnet(x,y,alpha=1)
lambda_1se <- cv.out$lambda.1se

x.test <- model.matrix(medv ~.,test)[,-1] # remove intercept
y.test <- test$medv

predict <- predict(lasso.model, s=lambda_1se, newx = x.test)
MSE <- mean((y.test - predict)^2)

coefs <- predict(lasso.model, type = 'coefficients', s=lambda_1se)
nonzero_coefs <- coefs[coefs[, 1] != 0, ]

# Print the test error and non-zero coefficient estimates
MSE
```

```
## [1] 19.06084
```

```
pander(nonzero_coefs)
```

Table 7: Table continues below

(Intercept)	crim	chas	nox	rm	dis	tax
19.45	-0.03226	2.821	-3.801	4.44	-0.4385	-0.0005855

ptratio	lstat
-0.7298	-0.5714

(b) **(10 marks)** (Best subset selection) Do the best subset selection with `BIC` and choose the best model. Report the values of coefficient estimates in the best model.

```
subset <- regsubsets(medv ~.,Boston, nvmax=10)
reg.summary <- summary(subset)
best_bic <- which.min(reg.summary$bic)
best_model_coeffs <- coef(subset,best_bic)
pander(best_model_coeffs)
```

Table 9: Table continues below

(Intercept)	crim	zn	chas	nox	rm	dis	rad
41.45	-0.1217	0.04619	2.872	-18.26	3.673	-1.516	0.2839

tax	ptratio	lstat
-0.01229	-0.931	-0.5465

(c) **(10 marks)**

Comparing the LASSO chosen model and the best subset selected model, which is the better model? Explain why?

The best subset model fits all combinations of predictors, and therefor is computationally expensive and not suitable for large datasets. Compared to best subset, LASSO regression is much less expensive, and can be used on large datasets.

LASSO adds the LASSO penalty to the SSE which is a bias predictor, meaning it hinders interpretation of the model co-efficients and therefor is unhelpful for inference problems.

Given the size of the dataset is small, the need for intepretatbility and predictive performance I believe the best subset selection model is best in this situation.

(d) **(10 marks)** How can you improve the fit of the best subset selected model?

```
# The best subset selected model includes the following predictors:
predictors <- names(best_model_coeffs)[-1]  # Exclude intercept
formula <- as.formula(paste("medv ~", paste(predictors, collapse = " + ")))

# Using best predictors from subset model found in 2c, fit a model with the best predictors.
best_subset_model <- gam(formula, data=Boston)

# Improved model, based on transformed best_subset_model
subset_improved <- gam(medv ~ crim + zn + chas + nox + s(rm) + log(dis) + log(rad) + tax + ptratio + s(
```

```r
# Create output of BIC comparison between best_subset_model and subset_improved.
Names_subset <- c('Original Subset', 'Improved Subset')
BIC_subset <- c(BIC(best_subset_model), BIC(subset_improved))
pander(data.frame(Names_subset, BIC_subset))
```

Names_subset	BIC_subset
Original Subset	3085
Improved Subset	2844

Including a spline function for predictors rm, and lstat as well as a log transformation for dis and rad significantly improved the BIC of the subset model as seen from the decrease in BIC from the subset, to improved_subset model.