STAA 552: HW 5

Matthew Stoebe

See Canvas Calendar for due date.
60 points total, 4 points per problem unless otherwise noted.
Content for all questions is from section 07 or earlier. Add or delete code chunks as needed.

Breast Cancer (Q1 - Q9)

In this group of questions, we will be fitting logistic regression models using training data. This data is available from Canvas as `BCTrain.csv`. This includes data from n = 512 subjects. This data is a subset of a larger data set that is publicly available from UCI Machine Learning Repository and Kaggle as "Breast Cancer Wisconsin (Diagnostic) Data Set". The columns include:

- id number (should NOT be used for model fitting)
- **diagnosis (0,1) should be used as the response for all analyses.** Malignant tumors are represented with a value of 1; benign tumors are represented with a value of 0.
- 10 predictor variables that are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

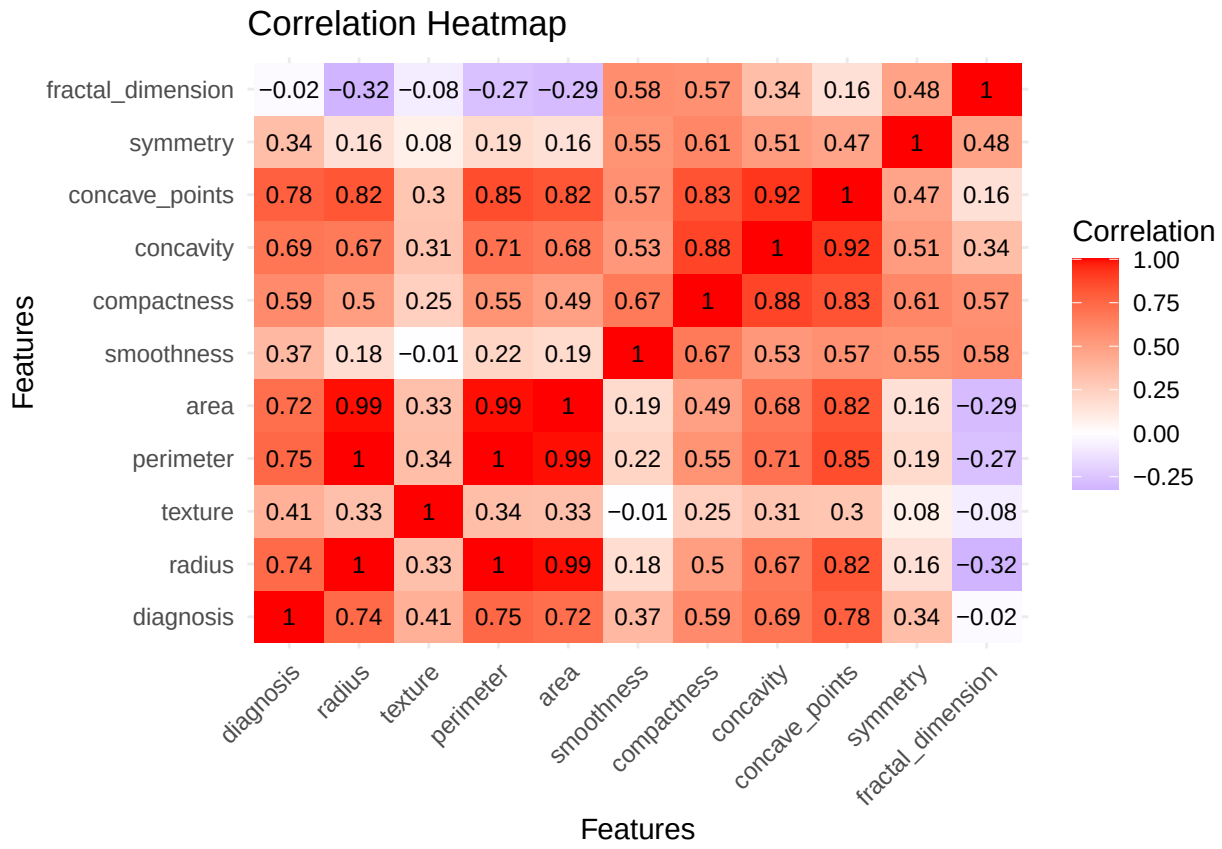
The goal is to predict diagnosis (Y) based on the predictor variables (X's).

Q1 (2 pts)

Calculate pairwise (Pearson) correlations between all predictors. Round your result to two decimal places.

```
##                  diagnosis radius texture perimeter  area smoothness
## diagnosis             1.00   0.74    0.41      0.75  0.72       0.37
## radius                0.74   1.00    0.33      1.00  0.99       0.18
## texture               0.41   0.33    1.00      0.34  0.33      -0.01
## perimeter             0.75   1.00    0.34      1.00  0.99       0.22
## area                  0.72   0.99    0.33      0.99  1.00       0.19
## smoothness            0.37   0.18   -0.01      0.22  0.19       1.00
## compactness           0.59   0.50    0.25      0.55  0.49       0.67
## concavity             0.69   0.67    0.31      0.71  0.68       0.53
## concave_points        0.78   0.82    0.30      0.85  0.82       0.57
## symmetry              0.34   0.16    0.08      0.19  0.16       0.55
## fractal_dimension    -0.02  -0.32   -0.08     -0.27 -0.29       0.58
##                  compactness concavity concave_points symmetry
## diagnosis               0.59      0.69           0.78     0.34
## radius                  0.50      0.67           0.82     0.16
```

```
## texture                  0.25      0.31           0.30      0.08
## perimeter                0.55      0.71           0.85      0.19
## area                     0.49      0.68           0.82      0.16
## smoothness               0.67      0.53           0.57      0.55
## compactness              1.00      0.88           0.83      0.61
## concavity                0.88      1.00           0.92      0.51
## concave_points           0.83      0.92           1.00      0.47
## symmetry                 0.61      0.51           0.47      1.00
## fractal_dimension        0.57      0.34           0.16      0.48
##                   fractal_dimension
## diagnosis                     -0.02
## radius                        -0.32
## texture                       -0.08
## perimeter                     -0.27
## area                          -0.29
## smoothness                     0.58
## compactness                    0.57
## concavity                      0.34
## concave_points                 0.16
## symmetry                       0.48
## fractal_dimension              1.00
```



*****

Q2

For all further questions, we will **DROP perimeter, area and concave_points**. Based on your results from the previous question, explain why this is reasonable. Be brief, but specific.

Response
These features are highly correlated and should be dropped to avoid multi-colinearity in the model ******

Q3

Perform model selection using backwards elimination. This should be done using p-values and with alpha = 0.05. Show the coefficients table for your final model. You do NOT need to show the intermediate output, only the final model.
NOTE: For purposes of predictive modeling, backwards elimination would NOT be a common choice. We are using this approach here for practice.

```
##
## Call:
## glm(formula = formula, family = binomial, data = model_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -42.39232    5.57005  -7.611 2.72e-14 ***
## radius        1.35299    0.17859   7.576 3.57e-14 ***
## texture       0.36479    0.06599   5.528 3.24e-08 ***
## smoothness   97.38729   22.62965   4.304 1.68e-05 ***
## concavity    16.02302    4.33692   3.695  0.00022 ***
## symmetry     25.08473   12.14620   2.065  0.03890 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 674.30  on 511  degrees of freedom
## Residual deviance: 137.75  on 506  degrees of freedom
## AIC: 149.75
##
## Number of Fisher Scoring iterations: 8
```

Q4

Perform model selection using **AIC** all subsets selection. Show the coefficients table for your final model. You do NOT need to show the intermediate output, only the final model.

```
##
## Call:
## glm(formula = diagnosis ~ radius + texture + smoothness + concavity +
##     symmetry + fractal_dimension, family = binomial, data = data)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -36.31179    6.55817  -5.537 3.08e-08 ***
## radius               1.19569    0.19855   6.022 1.72e-09 ***
## texture              0.36832    0.06618   5.565 2.62e-08 ***
## smoothness         113.28403   25.45504   4.450 8.57e-06 ***
## concavity           22.66536    6.22192   3.643  0.00027 ***
## symmetry            26.98767   12.33009   2.189  0.02861 *
## fractal_dimension -101.81291   66.42264  -1.533  0.12532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 674.30  on 511  degrees of freedom
## Residual deviance: 135.26  on 505  degrees of freedom
## AIC: 149.26
##
## Number of Fisher Scoring iterations: 8
```

Q5

Consider your results to the previous questions.

Q5A (2 pts)

Is it surprising that the selected models (from Q3 and Q4) are different? Briefly comment.

Response It is not surprising. Different methods yield different results, and the P value is rather simplistic as it does not consider the number of parameters in the model alongside the goodness of fit. AIC can favor a more complex model with a better fit. *****

Q5B (2pts)

Is it surprising that the selected model (from Q4) includes a predictor with p > 0.05? Briefly comment.

Response No. There are more complex things to consider than just p value in selecting parameters for a model. P Significance is not supreme in model building or parameter selection. You instead want to look at the incremental uplift you are getting with a parameter relative to other versions of the model. Again, AIC allows us to have a more robust and nuanced view in model selection that the P value elimination does not. *****

Important Note: For all further questions, we will use the AIC selected model from Q4.

Q6

Calculate McFadden's "pseudo" R2 value for this model.

R2 = 0.7994026

```
## 'log Lik.' 0.7994026 (df=7)
```

Q7

Run the Hosmer-Lemeshow test to test for lack of fit (using g = 10). Using alpha = 0.05, give a conclusion in context.

Response

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  final_model$y, fitted(final_model)
## X-squared = 9.6831, df = 8, p-value = 0.288
```

With a high p value we can conclude that our model is simply not that good. and is not statistically significant at a .05 level *****

Q8 (6 pts)

Suppose that a cancer is classified as malignant if $\hat{p} > 0.5$ and benign if $\hat{p} \leq 0.5$ Calculate the accuracy, TPR (true positive rate = sensitivity), TNR (true negative rate = specificity) for this model (based on the original, **training** data).

Accuracy: 0.9414062 TPR: 0.9392265 TNR: 0.9425982

```
## [1] 0.9414062
```

```
## [1] 0.9392265
```

```
## [1] 0.9425982
```

Q9 (6 ps)

Now use the **test** data (`BDTest.csv`) to calculate the accuracy, TPR (true positive rate = sensitivity), TNR (true negative rate = specificity) for this model.

Accuracy:
TPR:
TNR:

```
## [1] 0.8947368
```

```
## [1] 0.9047619
```

```
## [1] 0.8888889
```

Extinction (Q10 - Q16)

A study was conducted in the Krunnit Islands between 1949 and 1959. In each of N = 18 islands, the "Species at Risk" give the number of bird species present in 1949 (corresponding to n_i = number of "trials"). The "Extinctions" represent those not present in 1959 (corresponding to y_i = number of "events"). This is grouped data. See the file `Extinction.csv` on Canvas for the data.

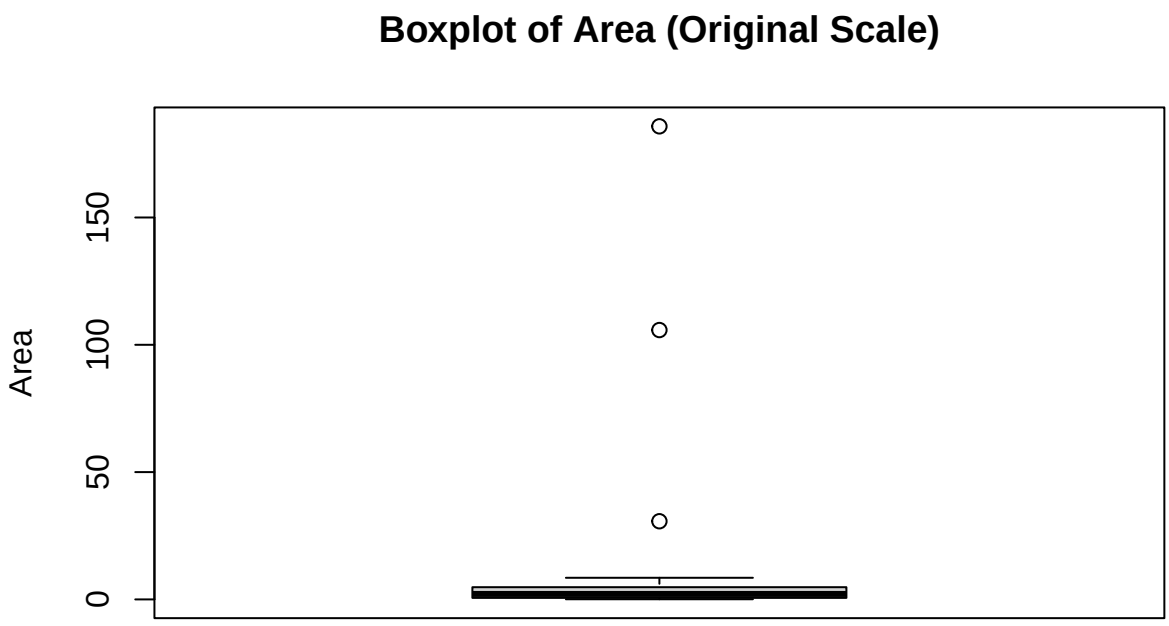
Our goal is to model probability of extinction as a function of island Area. Hence **extinction** is the event of interest for all questions.

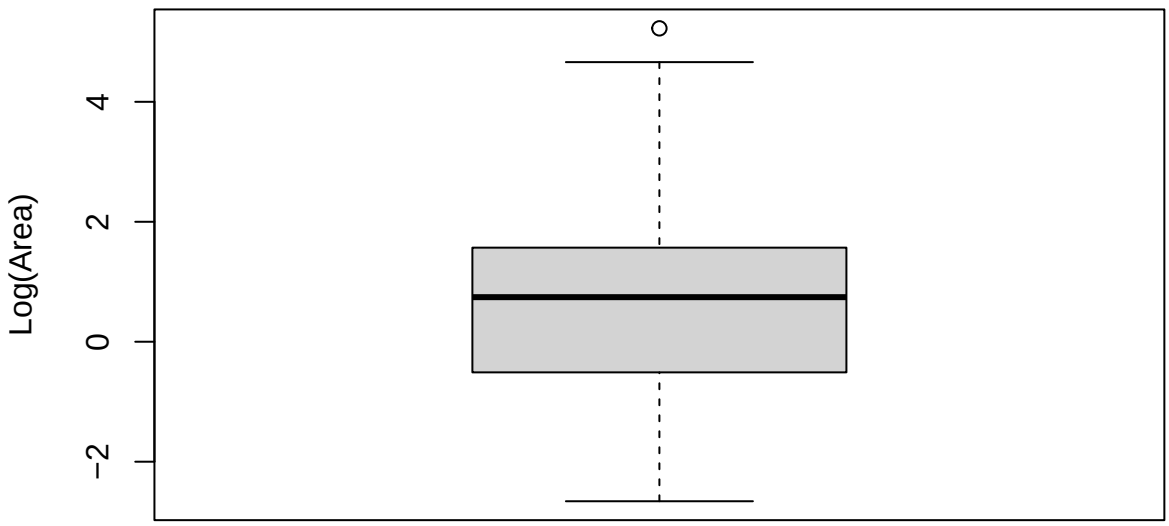
Q10

We will start with some simple exploratory graphs.

Q10A (2 pts)

Construct boxplots for Area on both the original and log transformed scales.

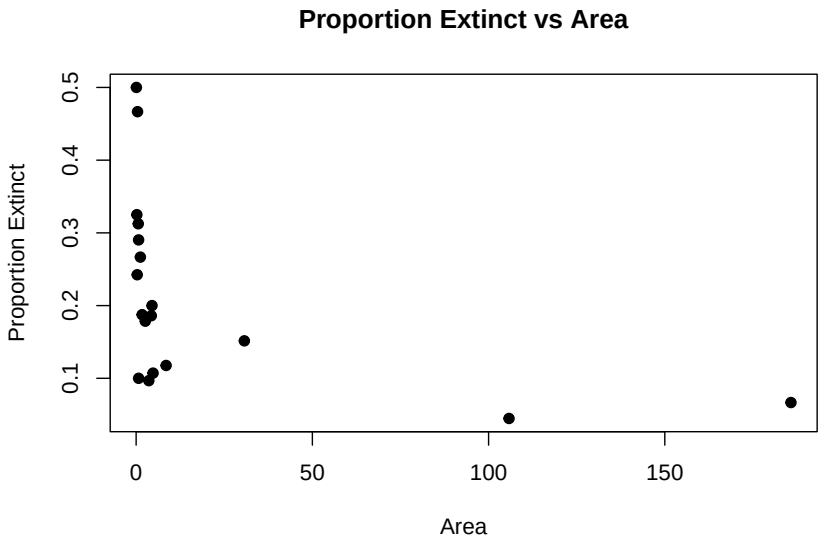

Boxplot of Area (Original Scale)
Area
0
50
100
150




```
##           Island  Area SpeciesAtRisk Extinctions
## 1     Ulkokrunni 185.8            75           5
## 2      Maakrunni 105.8            67           3
## 3      Ristikari  30.7            66          10
## 4 Isonkivenletto   8.5            51           6
## 5 Heitakraasukka   4.8            28           3
## 6      Kraasukka   4.5            20           4
```

Q10B (2 pts)

Construct a scatter plot of proportion extinct vs Area.



Extinction Model 1 (Q11 - Q12)

For this group of questions, we consider logistic regression with **Area** as the predictor.

Q11 (2 pts)

Fit the model and show the `summary()` output.

```
##
## Call:
## glm(formula = cbind(Extinctions, SpeciesAtRisk - Extinctions) ~
##     Area, family = binomial, data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.305957   0.117339 -11.130  < 2e-16 ***
## Area        -0.010121   0.002684  -3.771 0.000163 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45.338  on 17  degrees of freedom
## Residual deviance: 24.661  on 16  degrees of freedom
## AIC: 87.993
##
## Number of Fisher Scoring iterations: 4
```

Q12

Conduct a deviance-based lack-of-fit test for Model 1. Give the test statistic, df and p-value. **Using alpha = 0.10**, give a conclusion in context.

X2 = 24.66 df = 16
p = .07
Conclusion in context: Since the p-value is less than alpha = 0.10, we reject the null hypothesis. This suggests that there is evidence of lack of fit for Model 2, and the model does not adequately fit the data.

```
## [1] 24.66058
```

```
## [1] 16
```

```
## [1] 0.07603674
```

Extinction Model 2 (Q13 - Q15)

For this group of questions, we consider logistic regression with **log(Area)** as the predictor.

Q13 (2 pts)

Fit the model and show the `summary()` output.

```
##
## Call:
## glm(formula = cbind(Extinctions, SpeciesAtRisk - Extinctions) ~
##     log(Area), family = binomial, data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.19620    0.11845 -10.099  < 2e-16 ***
## log(Area)   -0.29710    0.05485  -5.416 6.08e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45.338  on 17  degrees of freedom
## Residual deviance: 12.062  on 16  degrees of freedom
## AIC: 75.394
##
## Number of Fisher Scoring iterations: 4
```

Q14

Conduct a deviance-based lack-of-fit test for Model 2. Give the test statistic, df and p-value. Using **alpha = 0.10**, give a conclusion in context.

X2 = 12.06 df = 16 p = .7397
Conclusion in context: Since the p-value is greater than alpha = 0.10, we fail to reject the null hypothesis. This suggests that there is no evidence of lack of fit for Model 2, and the model adequately fits the data.

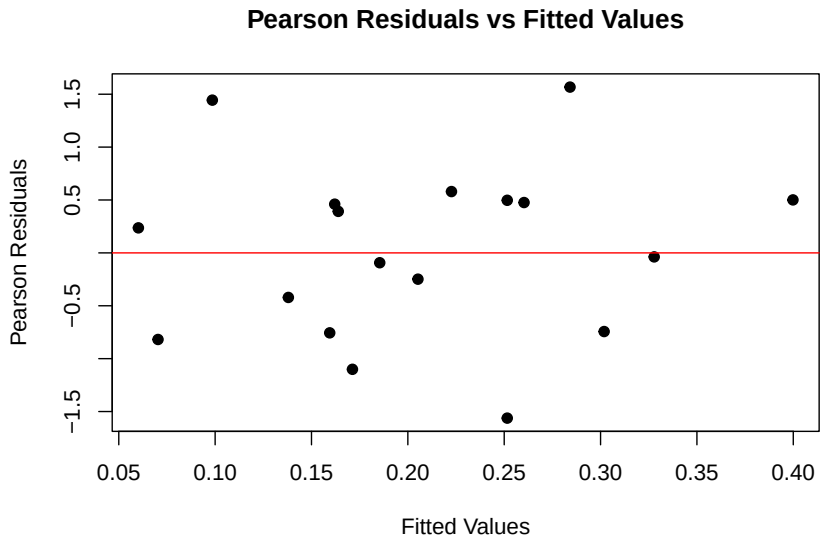
```
## [1] 12.06151
```

```
## [1] 16
```

```
## [1] 0.7397351
```

Q15

Using Model 2, create a plot of Pearson residuals versus fitted values. Do the residuals indicate any evidence of overdispersion?



Discussion There is no Clear over dispersion in these residuals/ *****

Q16 (2 pts)

Comparing Model 1 (using Area) and Model2 (using log(Area)), which model is preferred based on AIC?

Response

Model two has an AIC of 75 and Model one has an AIC of 87. With this in mind, we can select the second model as it is our preferred and more accurate choice.

Appendix

```
#Retain this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
library(ResourceSelection)
library(ggplot2)
library(reshape2)
library(MASS)

#Q1
data <- read.csv("Data/BCTrain.csv")

corr_data <- data[, -1]

coors <- round(cor(corr_data),2)
print(coors)

melted_coors <- melt(coors)

# Create heatmap
ggplot(data = melted_coors, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = value), color = "black", size = 3) +  # Add coefficients as text
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  labs(title = "Correlation Heatmap", x = "Features", y = "Features", fill = "Correlation")

data <- data[, !(names(data) %in% c("perimeter", "area", "concave_points", "id"))]

#Q3
#I initially used the step function here but it defaults to AIC. I dont see a function for P value so i

backward_elimination_p <- function(model, threshold = 0.05) {
```

```r
  model_data <- model$data

  while (TRUE) {
    #Get Maximum P value
    model_summary <- summary(model)
    p_values <- coef(model_summary)[, "Pr(>|z|)"]
    p_values <- p_values[-1]
    max_p <- max(p_values, na.rm = TRUE)

#Remove Column from maximum P Value if it is above threshold
    if (max_p > threshold) {
      var_to_remove <- names(p_values)[which.max(p_values)]
      formula <- as.formula(
        paste("diagnosis ~", paste(setdiff(names(p_values), var_to_remove), collapse = " + "))
      )
      model <- glm(formula, data = model_data, family = binomial)
    }
    else {
      break
    }
  }
  return(model)
}

full_model <- glm(diagnosis ~ ., data = data, family = binomial)
final_model <- backward_elimination_p(full_model)
summary(final_model)



#Q4
step_model <- step(full_model, direction = "both", trace = 0)

summary(step_model)
#Q6
final_model <- step_model

final_loglike <- logLik(final_model)
null_model <- glm(diagnosis ~ 1, data=data, family=binomial)
null_loglike <- logLik(null_model)

r2_ish <- 1- (final_loglike/null_loglike)
r2_ish

#Q7
hl_test <- hoslem.test(final_model$y, fitted(final_model),g=10)

hl_test
#Q8
pred_probs <- predict(final_model, type = "response")

pred_class <- ifelse(pred_probs > 0.5, 1, 0)
```

```r
conf_matrix <- table(Predicted = pred_class, Actual = data$diagnosis)

# Calculate accuracy, TPR, and TNR
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
tpr <- conf_matrix[2, 2] / sum(conf_matrix[2, ])
tnr <- conf_matrix[1, 1] / sum(conf_matrix[1, ])

accuracy
tpr
tnr

#Q9
# Q9
#Get Test Data
test_data <- read.csv("Data/BCTest.csv")

test_pred_probs <- predict(final_model, newdata = test_data, type = "response")
test_pred_class <- ifelse(test_pred_probs > 0.5, 1, 0)

# Get Metrics
test_conf_matrix <- table(Predicted = test_pred_class, Actual = test_data$diagnosis)

test_accuracy <- sum(diag(test_conf_matrix)) / sum(test_conf_matrix)
test_tpr <- test_conf_matrix[2, 2] / sum(test_conf_matrix[2, ])
test_tnr <- test_conf_matrix[1, 1] / sum(test_conf_matrix[1, ])

test_accuracy
test_tpr
test_tnr

#Q10
data <- read.csv("Data/Extinction.csv")

# Boxplot for original Area
boxplot(data$Area, main = "Boxplot of Area (Original Scale)", ylab = "Area")

# Boxplot for log-transformed Area
boxplot(log(data$Area), main = "Boxplot of Area (Log Transformed)", ylab = "Log(Area)")

head(data)
#Q11
data$proportion_extinct <- data$Extinctions / data$SpeciesAtRisk

# Scatter plot of proportion extinct vs Area
plot(data$Area, data$proportion_extinct, main = "Proportion Extinct vs Area", xlab = "Area", ylab = "Pr
#Q11
model1 <- glm(cbind(Extinctions, SpeciesAtRisk - Extinctions) ~ Area, family = binomial, data = data)
summary(model1)

#Q12
deviance_statistic <- model1$deviance
df <- model1$df.residual
p_value <- 1 - pchisq(deviance_statistic, df)
```

```r
deviance_statistic
df
p_value
#Q13
model2 <- glm(cbind(Extinctions, SpeciesAtRisk - Extinctions) ~ log(Area), family = binomial, data = da

summary(model2)
#Q14
deviance_statistic <- model2$deviance
df <- model2$df.residual
p_value <- 1 - pchisq(deviance_statistic, df)

deviance_statistic
df
p_value
#Q15
pearson_resid <- residuals(model2, type = "pearson")
fitted_values <- fitted(model2)
plot(fitted_values, pearson_resid, xlab = "Fitted Values", ylab = "Pearson Residuals", main = "Pearson
abline(h = 0, col = "red")
```