Project

Sagnik Chakravarty

```
library(readr)
library(dplyr)
library(ggplot2)
library(knitr)
library(tidyverse)
library(car)
library(ResourceSelection)
library(pROC)
library(caret)
library(MASS)
library(pacman)
data <- read_csv(file = 'chd.csv')
data$famhist<- factor(data$famhist,
                      levels = c("Absent", "Present"),
                      labels = c(0, 1))
data$chd <- factor(data$chd)
kable(head(data), format = 'latex')
```

sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
128	0.0	2.63	23.88	0	45	21.59	6.54	57	0
160	3.0	9.19	26.47	1	39	28.25	14.40	54	1
162	7.0	7.67	34.34	1	33	30.77	0.00	62	0
136	5.8	5.90	27.55	0	65	25.71	14.40	59	0
170	0.4	4.11	42.06	1	56	33.10	2.06	57	0
130	4.0	2.40	17.42	0	60	22.05	0.00	40	0

```
str(data)
```

```
spc_tbl_ [420 x 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ sbp      : num [1:420] 128 160 162 136 170 130 150 122 132 132 ...
 $ tobacco  : num [1:420] 0 3 7 5.8 0.4 4 0.18 4.18 2.8 7.2 ...
 $ ldl      : num [1:420] 2.63 9.19 7.67 5.9 4.11 2.4 4.14 9.05 4.79 3.65 ...
 $ adiposity: num [1:420] 23.9 26.5 34.3 27.6 42.1 ...
 $ famhist  : Factor w/ 2 levels "0","1": 1 2 2 1 2 1 1 2 2 2 ...
 $ typea    : num [1:420] 45 39 33 65 56 60 53 44 50 56 ...
 $ obesity  : num [1:420] 21.6 28.2 30.8 25.7 33.1 ...
 $ alcohol  : num [1:420] 6.54 14.4 0 14.4 2.06 ...
 $ age      : num [1:420] 57 54 62 59 57 40 44 52 48 34 ...
 $ chd      : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 2 1 1 ...
 - attr(*, "spec")=
  .. cols(
  ..   sbp = col_double(),
  ..   tobacco = col_double(),
  ..   ldl = col_double(),
```

```
..   adiposity = col_double(),
..   famhist = col_character(),
..   typea = col_double(),
..   obesity = col_double(),
..   alcohol = col_double(),
..   age = col_double(),
..   chd = col_double()
.. )
- attr(*, "problems")=<externalptr>
```

EDA

```
# CHD prevalence
kable(table(data$chd), format = 'latex')
```

Var1	Freq
0	276
1	144

```
kable(prop.table(table(data$chd)), format = 'latex')
```

Var1	Freq
0	0.6571429
1	0.3428571

```
# Continuous variables summary
continuous_vars <- c("sbp",
                     "tobacco",
                     "ldl",
                     "adiposity",
                     "obesity",
                     "alcohol",
                     "age",
                     "typea")
kable(sapply(data[continuous_vars],
             function(x) c(Mean = mean(x),
                           SD = sd(x),
                           Min = min(x),
                           Max = max(x))),
      format = 'latex')
```

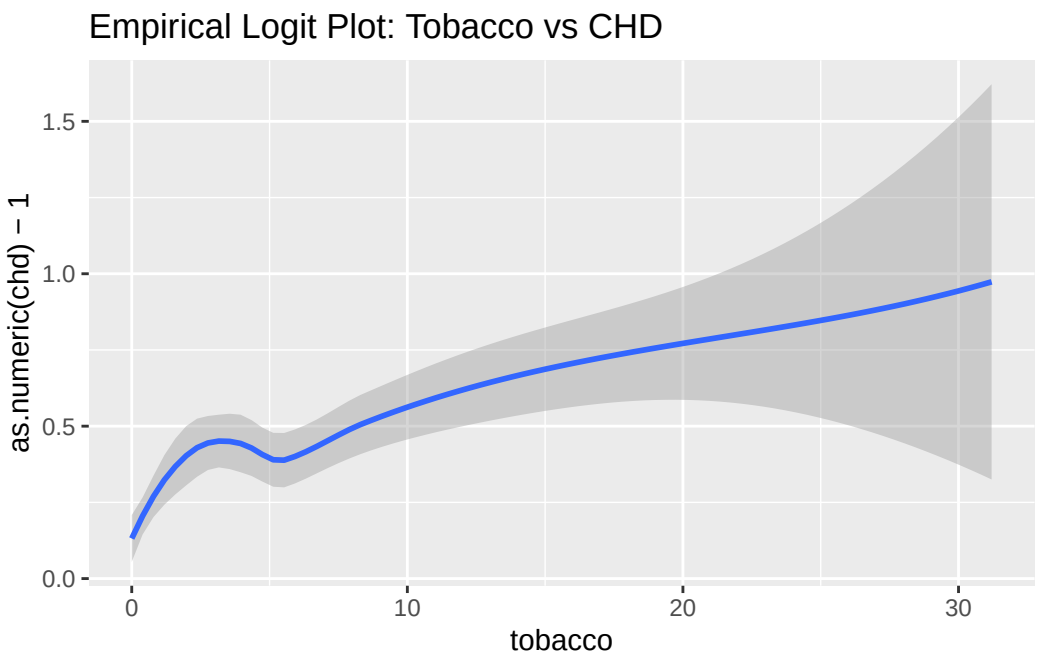
	sbp	tobacco	ldl	adiposity	obesity	alcohol	age	typea
Mean	138.49286	3.734809	4.725786	25.461000	26.070024	17.17221	43.07381	52.971429
SD	20.51642	4.688402	2.069873	7.763727	4.261565	24.61094	14.52351	9.665733
Min	101.00000	0.000000	0.980000	6.740000	14.700000	0.00000	15.00000	13.000000
Max	218.00000	31.200000	15.330000	42.490000	46.580000	147.19000	64.00000	78.000000

```
# Categorical variable summary
kable(table(data$famhist), format = 'latex')
```

Var1	Freq
0	243
1	177

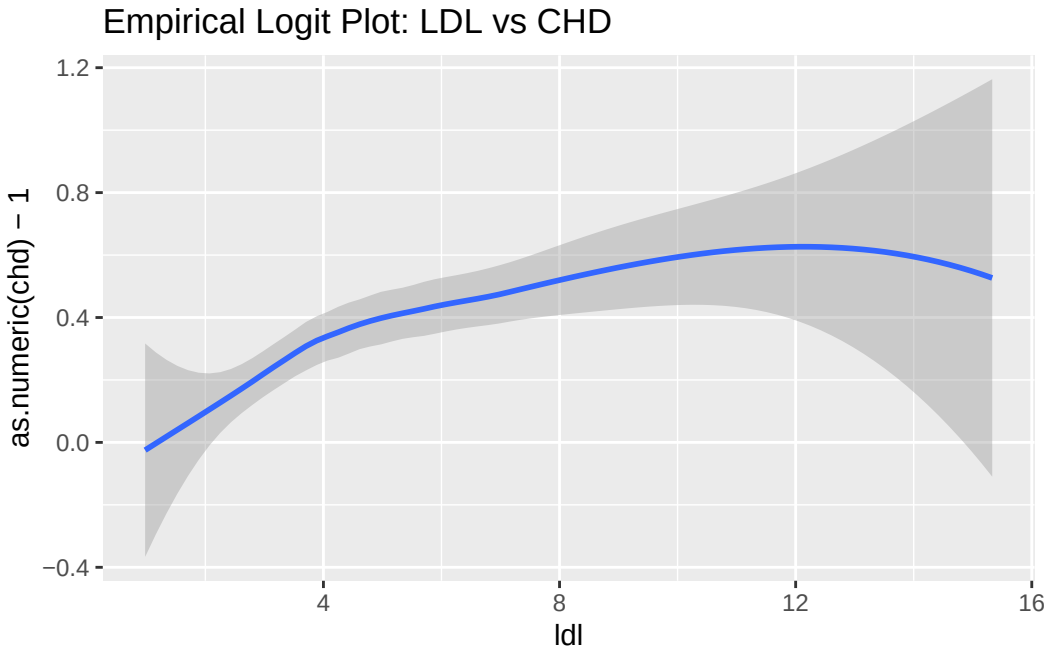
```r
# Bivariate relationships
# Empirical logit plots for continuous variables
ggplot(data, aes(x = tobacco, y = as.numeric(chd)-1)) +
  geom_smooth(method = "loess") +
  labs(title = "Empirical Logit Plot: Tobacco vs CHD")
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```r
ggplot(data, aes(x = ldl, y = as.numeric(chd)-1)) +
  geom_smooth(method = "loess") +
  labs(title = "Empirical Logit Plot: LDL vs CHD")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```r
# Full logistic regression model
full_model <- glm(chd ~ sbp + tobacco + ldl + adiposity + famhist +
                  typea + obesity + alcohol + age,
                family = binomial(link = "logit"), data = data)

# Model summary
summary(full_model)
```

```
Call:
glm(formula = chd ~ sbp + tobacco + ldl + adiposity + famhist +
    typea + obesity + alcohol + age, family = binomial(link = "logit"),
    data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.979369   1.369834  -4.365 1.27e-05 ***
sbp          0.007746   0.005958   1.300 0.193526
tobacco      0.081893   0.027356   2.994 0.002758 **
ldl          0.178967   0.062625   2.858 0.004267 **
adiposity    0.014893   0.030536   0.488 0.625744
famhist1     0.920423   0.239991   3.835 0.000125 ***
typea        0.030187   0.012873   2.345 0.019026 *
obesity     -0.055503   0.045343  -1.224 0.220925
alcohol      0.001774   0.004672   0.380 0.704129
age          0.044239   0.012707   3.482 0.000498 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 540.05  on 419  degrees of freedom
Residual deviance: 428.74  on 410  degrees of freedom
AIC: 448.74

Number of Fisher Scoring iterations: 5
```

```r
# Likelihood ratio test
null_model <- glm(chd ~ 1, family = binomial, data = data)
lmtest::lrtest(null_model, full_model)
```

```
Likelihood ratio test

Model 1: chd ~ 1
Model 2: chd ~ sbp + tobacco + ldl + adiposity + famhist + typea + obesity +
    alcohol + age
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   1 -270.02
2  10 -214.37  9 111.31  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# McFadden's pseudo R-squared
1 - (logLik(full_model)/logLik(null_model))
```

```
'log Lik.' 0.2061132 (df=10)
```

```r
# Stepwise model selection
step_model <- stepAIC(full_model, direction = "both", trace = FALSE)
summary(step_model)
```

```
Call:
glm(formula = chd ~ tobacco + ldl + famhist + typea + age, family = binomial(link = "logit"),
    data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.98647    0.95421  -6.274 3.52e-10 ***
tobacco      0.08444    0.02665   3.168 0.001534 **
ldl          0.16578    0.05770   2.873 0.004064 **
famhist1     0.91050    0.23743   3.835 0.000126 ***
typea        0.02794    0.01272   2.197 0.028023 *
age          0.04939    0.01079   4.577 4.72e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 540.05  on 419  degrees of freedom
Residual deviance: 432.37  on 414  degrees of freedom
AIC: 444.37

Number of Fisher Scoring iterations: 5
```

```r
# Final model (based on analysis in report)
final_model <- full_model
```

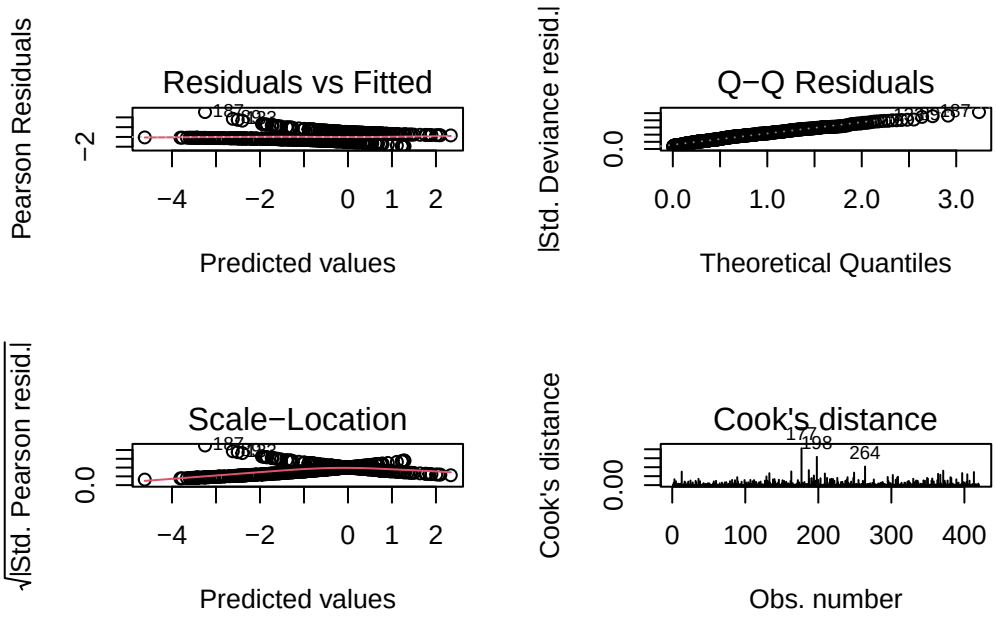
```r
# Variance Inflation Factors (check multicollinearity)
vif(final_model)
```

```
     sbp  tobacco      ldl adiposity  famhist    typea  obesity  alcohol
1.139224 1.156813 1.187193 3.326718 1.028312 1.069234 2.507246 1.067645
     age
1.647164
```

```r
# Hosmer-Lemeshow test
hoslem.test(final_model$y, fitted(final_model), g = 10)
```

```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  final_model$y, fitted(final_model)
X-squared = 3.0373, df = 8, p-value = 0.932
```

```r
# Residual analysis
par(mfrow = c(2,2))
plot(final_model, which = 1:4)
```



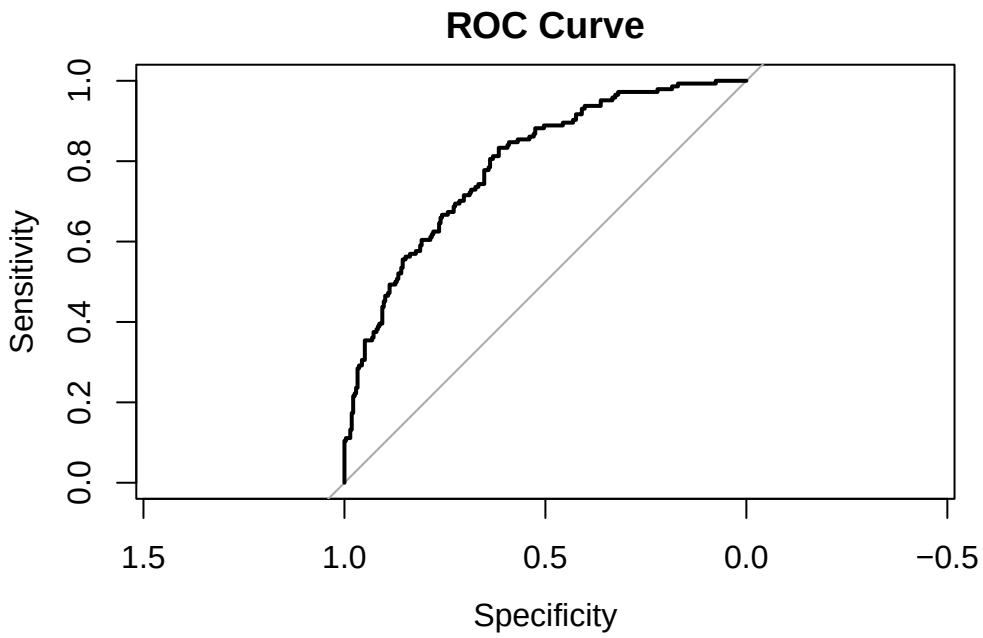

```r
par(mfrow = c(1,1))
```

```r
################################
# Predictive Accuracy Assessment #
################################

# Predicted probabilities
data$pred_prob <- predict(final_model, type = "response")

# Confusion matrix (0.5 cutoff)
# Convert both vectors to factors with EXPLICIT LEVELS
predicted <- factor(ifelse(data$pred_prob > 0.5, "Yes", "No"),
                    levels = c("No", "Yes"))
actual <- factor(data$chd, levels = c("No", "Yes"))

# Create confusion matrix with aligned factors
conf_mat <- confusionMatrix(predicted, actual)

# ROC curve analysis
roc_obj <- roc(data$chd, data$pred_prob)
```

```
Setting levels: control = 0, case = 1
```

```
Setting direction: controls < cases
```

```r
plot(roc_obj, main = "ROC Curve")
```




```
auc(roc_obj)
```

```
Area under the curve: 0.7932
```

```
# Optimal cutoff using Youden's index
coords(roc_obj, "best", best.method = "youden")
```

```
  threshold specificity sensitivity
1 0.2888442    0.615942   0.8333333
```

```
########################
# Results Interpretation #
########################

# Odds ratios and CI
odds_ratios <- exp(coef(final_model))
ci <- exp(confint(final_model))
```

```
Waiting for profiling to be done...
```

```
results_table <- data.frame(
  Predictor = names(odds_ratios),
  OR = round(odds_ratios, 3),
  CI_Lower = round(ci[,1], 3),
  CI_Upper = round(ci[,2], 3)
)

print(results_table)
```

```
                Predictor    OR CI_Lower CI_Upper
(Intercept) (Intercept) 0.003    0.000    0.035
```

```
sbp                 sbp 1.008    0.996    1.020
tobacco         tobacco 1.085    1.030    1.147
ldl                 ldl 1.196    1.061    1.357
adiposity     adiposity 1.015    0.956    1.078
famhist1       famhist1 2.510    1.573    4.036
typea             typea 1.031    1.005    1.058
obesity         obesity 0.946    0.863    1.032
alcohol         alcohol 1.002    0.993    1.011
age                 age 1.045    1.020    1.072
```

```
# Clinical interpretation
cat("Key Interpretation:\n")
```

```
Key Interpretation:
```

```
cat("- Each additional cigarette/day increases CHD odds by", round(100*(results_table["tobacco"
```

```
- Each additional cigarette/day increases CHD odds by 8.5 %
```

```
cat("- Each unit increase in LDL increases CHD odds by", round(100*(results_table["ldl","OR"]-1
```

```
- Each unit increase in LDL increases CHD odds by 19.6 %
```

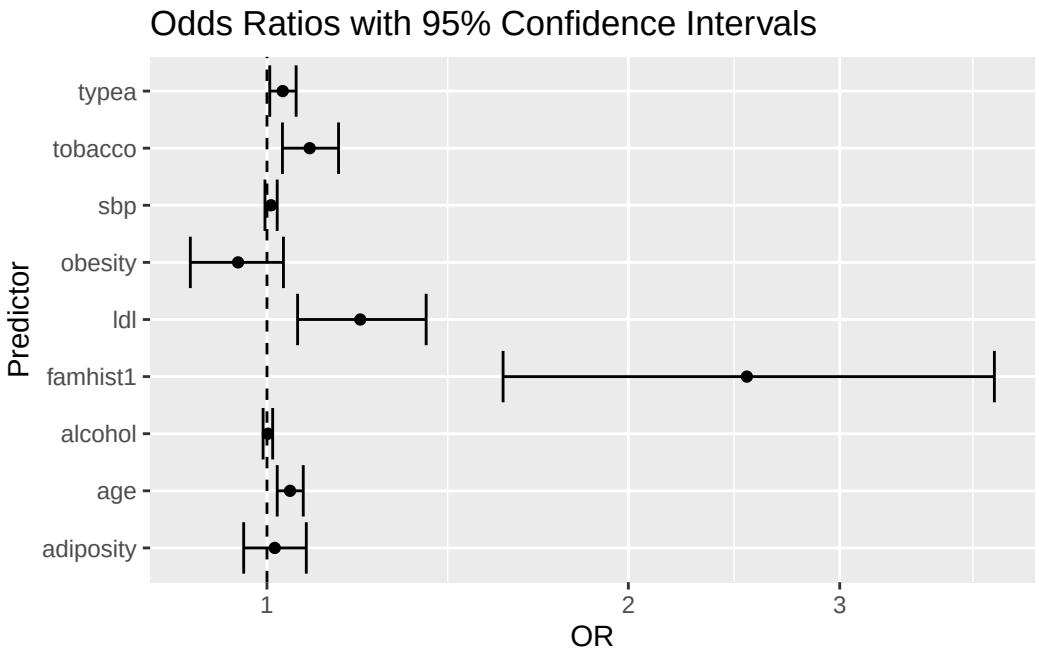
```
cat("- Family history increases CHD odds by", round(results_table["famhistYes","OR"],1), "times
```

```
- Family history increases CHD odds by NA times
```

```
########################
# Visualization Code #
########################

# Coefficient plot
coef_plot <- data.frame(Predictor = names(coef(final_model))[-1],
                        OR = exp(coef(final_model))[-1],
                        CI_Lower = exp(confint(final_model))[-1,1],
                        CI_Upper = exp(confint(final_model))[-1,2])
```

```
Waiting for profiling to be done...
Waiting for profiling to be done...
```

```
ggplot(coef_plot, aes(x = OR, y = Predictor)) +
  geom_point() +
  geom_errorbarh(aes(xmin = CI_Lower, xmax = CI_Upper)) +
  geom_vline(xintercept = 1, linetype = "dashed") +
  scale_x_log10() +
  labs(title = "Odds Ratios with 95% Confidence Intervals")
```




```r
# Predicted probability distribution
ggplot(data, aes(x = pred_prob, fill = chd)) +
  geom_density(alpha = 0.5) +
  labs(x = "Predicted Probability of CHD",
       title = "Probability Distribution by CHD Status")
```