Example R Markdown

0.1 Data

In this report, we use the `NHANESsample` dataset from the `HDSinRdata` package, which comes from NHANES 1999-2018 and was downloaded from the **nhanesA** package. This dataset includes information on lead levels, blood pressure, and demographic variables for 31,265 subjects.

0.2 Exploratory Analysis and Model Fitting

We first created a subset of the `NHANESsample` dataset that contains only the age, sex, income, smoking status, lead level, hypertension status, and alcohol use variables. Then, we fit a logistic regression with hypertension as the outcome and include the main effects of all 6 covariates. In this model, only age, sex, income, and alcohol use were significantly associated with hypertension, so we fit a new model containing just these independent variables and then conducted a forward stepwise selection procedure based on AIC to find possible interactions that could improve the model's fit. Our final model included interactions between sex and all three other covariates as well as an interaction between age and income. The coefficients from this model can be seen in Table 1 below, and the resulting ROC curve is shown in Figure 1.

Table 1: Final Model Intercepts and Odds Ratios

Term	Estimate	Std. Error	P Value
(Intercept)	0.1424081	0.0991462	0.0000000
Age	1.0508650	0.0017498	0.0000000
Female	0.2655810	0.1114930	0.0000000
Income	1.0625008	0.0267370	0.0233615
Alcohol	0.9781570	0.0472717	0.6403598
Age:Female	1.0269481	0.0017346	0.0000000
Income:Female	0.9109710	0.0161170	0.0000000
Income:Age	0.9985685	0.0005254	0.0063996
Alcohol:Female	0.8927137	0.0687780	0.0989260

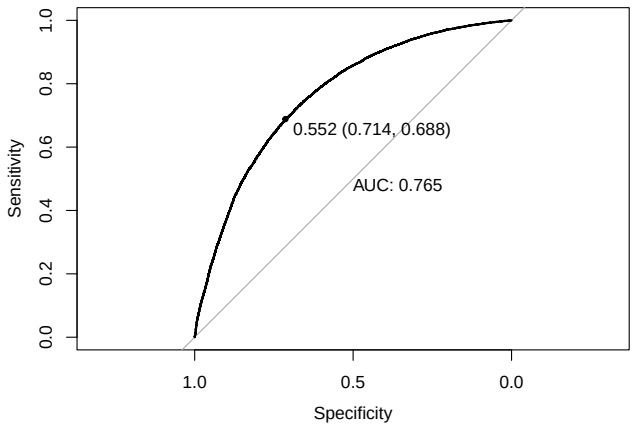

Figure 1: ROC Curve for Final Model

We can further examine the relationship between the two continuous predictors in our model and our outcome of interest by drawing our attention to the plots in Figure 2 below.

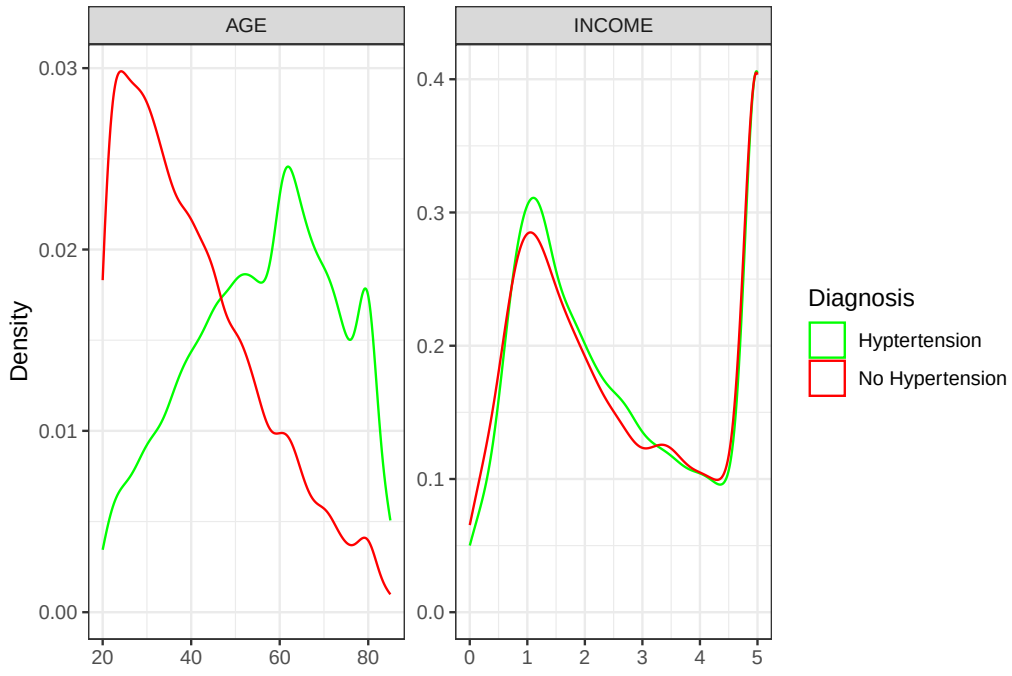

Figure 2: Predictor Variable Distributions by Diagnosis

1 Code Appendix

```
knitr::opts_chunk$set(message=FALSE,
                      warning=FALSE,
                      error=FALSE,
                      echo = FALSE,
                      fig.pos = "H",
                      out.extra = '')
library(tidyverse)
library(knitr)
library(kableExtra)
library(broom)
library(HDSinRdata)
library(pROC)
#load in data
data(NHANESsample)
#subset the data and clean
nhanes_sub <- NHANESsample %>%
  select(AGE, SEX, INCOME, SMOKE, LEAD, HYP, ALC)

#full model
original_mod <- glm(HYP ~ ., data = nhanes_sub, family = binomial)

#include only the significant main effects from the previous model
simple_mod <- glm(HYP ~ AGE + SEX + INCOME + ALC,
                  data = nhanes_sub, family = binomial)

#specify our upper bound for the model scope
fullmod <- glm(HYP ~ (AGE + SEX + INCOME + ALC)^2,
               data = nhanes_sub, family = binomial)

#perform forward stepwise selection
step_mod <- step(simple_mod, scope = list(upper = fullmod),
                 direction = "forward", trace = 0)

#view the summary
final_mod <- glm(HYP ~ AGE + SEX + INCOME + ALC + AGE:SEX + SEX:INCOME +
                   AGE:INCOME + SEX:ALC,
                 data = nhanes_sub, family = binomial)

tidy(final_mod, exponentiate=TRUE) %>%
  select(-statistic) %>%
  mutate(term = c("(Intercept)", "Age", "Female", "Income", "Alcohol",
                  "Age:Female", "Income:Female", "Income:Age",
                  "Alcohol:Female")) %>%
  kable(caption = "Final Model Intercepts and Odds Ratios",
        col.names = c("Term", "Estimate", "Std. Error", "P Value")) %>%
  kable_styling(latex_options = c("HOLD_position"),
                font_size=8)
roccurve <- roc(predictor=predict(final_mod, type="response"),
                    response=as.factor(final_mod$y),
                    levels = c(0,1), direction = "<")
plot(roccurve, print.auc=TRUE, print.thres = TRUE)
```

```r
nhanes_long <- nhanes_sub %>% select(HYP, AGE, INCOME) %>%
  pivot_longer(cols = AGE:INCOME) %>%
  mutate(hypertension_status = case_when(HYP == 1 ~ "Hyptertension",
                                         HYP == 0 ~ "No Hypertension"))

ggplot(data = nhanes_long) +
  geom_density(aes(x = value, y = ..density.., color = hypertension_status)) +
  facet_wrap(~ name, scales = "free") +
  labs(y = "Density", x = "") +
  scale_color_manual(name = "Diagnosis", values = c("green", "red")) +
  theme_bw()
```