STAT 650-Final Project

Safiya Geelani

10/02/2022

```
library(tidyverse)
library(ggplot2)
library(GGally)
library(janitor)
library(ggpubr)
```

Read in the data set and basic descriptive statistics

```
diabetes_data <- read_csv("diabetes.csv")
glimpse(diabetes_data)
summary(diabetes_data)
```

Cleaning the data

```
diabetes <- diabetes_data %>%
  mutate_at(vars(Outcome), as.factor) %>%
  rename(DiabetesPercent = DiabetesPedigreeFunction) %>%
  clean_names()
```

Exploratory Data Analysis/ Visualizing

```
# Histogram for numeric variables
diabetes %>% select_if(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~ key, scales = "free")

# Barplot for categorical variables
diabetes %>%
  ggplot(aes(fct_rev(fct_infreq(outcome)))) +
  geom_bar() + xlab("outcome")
```

```
ggpairs(diabetes[,c(9,1:4)])
ggpairs(diabetes[,c(9,5:8)])
```

Relationship between different variables

```
diabetes %>%
  ggplot() +
  geom_histogram(bins = 20, aes(x = bmi, fill = outcome))

diabetes %>%
  ggplot() +
  geom_histogram(bins = 20, aes(x = glucose, fill = outcome))

diabetes %>%
  ggplot(aes(x = age, y = pregnancies,
             color = outcome)) +
  geom_point() +
  geom_smooth(se = F)

diabetes %>%
  ggplot(aes(x = age, y = glucose,
             color = outcome)) +
  geom_point() +
  geom_smooth(se = F)

diabetes_smry <- diabetes %>%
group_by(outcome) %>%
summarise(count = n(),
          bmi = mean(bmi))
ggplot(diabetes_smry, aes(bmi, fct_reorder(outcome, bmi))) +
geom_point() + ylab("outcome")

diabetes_smry <- diabetes %>%
group_by(outcome) %>%
summarise(count = n(),
          insulin = mean(insulin))
ggplot(diabetes_smry, aes(insulin, fct_reorder(outcome, insulin))) +
geom_point() + ylab("outcome")
```

Variable selection

```
# Considering full model first
diabetes_full <- glm(outcome ~ ., family = "binomial", data = diabetes)
summary(diabetes_full)

# Removing one variable at a time
glm1 <- update(diabetes_full, ~ . -skin_thickness, data = diabetes)
```

```
summary(glm1)
glm2 <- update(glm1, ~ . -insulin, data = diabetes)
summary(glm2)
glm3 <- update(glm2, ~ . -age, data = diabetes)
summary(glm3)

#BIC stepwise selection
n <- nrow(diabetes)
diabetes_red <- step(diabetes_full, k = log(n))
summary(diabetes_red)
```

Making data frames for doing some predictions

```
# Effect of pregnancies holding other variables constant
diabetes1 <- with(diabetes, data.frame(pregnancies = seq(min(pregnancies),
                                                         max(pregnancies)),
                                       glucose = mean(glucose),
                                       blood_pressure = mean(blood_pressure),
                                       bmi = mean(bmi),
                                       diabetes_percent = mean(diabetes_percent)))

# Effect of glucose holding other variables constant
diabetes2 <- with(diabetes, data.frame(pregnancies = mean(pregnancies),
                                       glucose = seq(min(glucose),
                                                     max(glucose)),
                                       blood_pressure = mean(blood_pressure),
                                       bmi = mean(bmi),
                                       diabetes_percent = mean(diabetes_percent)))

# Effect of blood pressure holding other variables constant
diabetes3 <- with(diabetes, data.frame(pregnancies = mean(pregnancies),
                                       glucose = mean(glucose),
                                       blood_pressure = seq(min(blood_pressure),
                                                            max(blood_pressure)),
                                       bmi = mean(bmi),
                                       diabetes_percent = mean(diabetes_percent)))

# Effect of bmi holding other variables constant
diabetes4 <- with(diabetes, data.frame(pregnancies = mean(pregnancies),
                                       glucose = mean(glucose),
                                       blood_pressure = mean(blood_pressure),
                                       bmi = seq(min(bmi), max(bmi)),
                                       diabetes_percent = mean(diabetes_percent)))

# Effect of diabetes percentage holding other variables constant
diabetes5 <- with(diabetes, data.frame(pregnancies = mean(pregnancies),
                                       glucose = mean(glucose),
                                       blood_pressure = mean(blood_pressure),
                                       bmi = mean(bmi),
                                       diabetes_percent = seq(min(diabetes_percent),
                                                              max(diabetes_percent))))
```

Making predictions

```
diabetes_1 <- cbind(diabetes1, predict(glm3, newdata = diabetes1, type = "link", se = TRUE))
diabetes_1 <- within(diabetes_1, {
  pred_prob <- plogis(fit)
  lower <- plogis(fit - (1.96 * se.fit))
  upper <- plogis(fit + (1.96 * se.fit))})

diabetes_2 <- cbind(diabetes2, predict(glm3, newdata = diabetes2, type = "link", se = TRUE))
diabetes_2 <- within(diabetes_2, {
  pred_prob <- plogis(fit)
  lower <- plogis(fit - (1.96 * se.fit))
  upper <- plogis(fit + (1.96 * se.fit))})

diabetes_3 <- cbind(diabetes3, predict(glm3, newdata = diabetes3, type = "link", se = TRUE))
diabetes_3 <- within(diabetes_3, {
  pred_prob <- plogis(fit)
  lower <- plogis(fit - (1.96 * se.fit))
  upper <- plogis(fit + (1.96 * se.fit))})

diabetes_4 <- cbind(diabetes4, predict(glm3, newdata = diabetes4, type = "link", se = TRUE))
diabetes_4 <- within(diabetes_4, {
  pred_prob <- plogis(fit)
  lower <- plogis(fit - (1.96 * se.fit))
  upper <- plogis(fit + (1.96 * se.fit))})

diabetes_5 <- cbind(diabetes5, predict(glm3, newdata = diabetes5, type = "link", se = TRUE))
diabetes_5 <- within(diabetes_5, {
  pred_prob <- plogis(fit)
  lower <- plogis(fit - (1.96 * se.fit))
  upper <- plogis(fit + (1.96 * se.fit))})
```

Visualization of the predictions

```
a <- ggplot(diabetes_1, aes(x = pregnancies, y = pred_prob)) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.10) +
  geom_line(color = "coral") + theme_light() +
  ylab("Diabetes_pp") #pp means predicted probabilities

b <- ggplot(diabetes_2, aes(x = glucose, y = pred_prob)) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.10) +
  geom_line(color = "brown") +  theme_light() +
  ylab("Diabetes_pp")

c <- ggplot(diabetes_3, aes(x = blood_pressure, y = pred_prob)) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.10) +
  geom_line(color = "light green") +  theme_light() +
  ylab("Diabetes_pp")

d <- ggplot(diabetes_4, aes(x = bmi, y = pred_prob)) +
```

```
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.10) +
  geom_line(color = "sky blue") +  theme_light() +
  ylab("Diabetes_pp")

e <- ggplot(diabetes_5, aes(x = diabetes_percent, y = pred_prob)) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.10) +
  geom_line(color = "violet") +  theme_light() +
  ylab("Diabetes_pp")

ggarrange(a, b, c, d, e, 0, ncol = 2, nrow = 3)
```