Environmental Statistics Homework 3

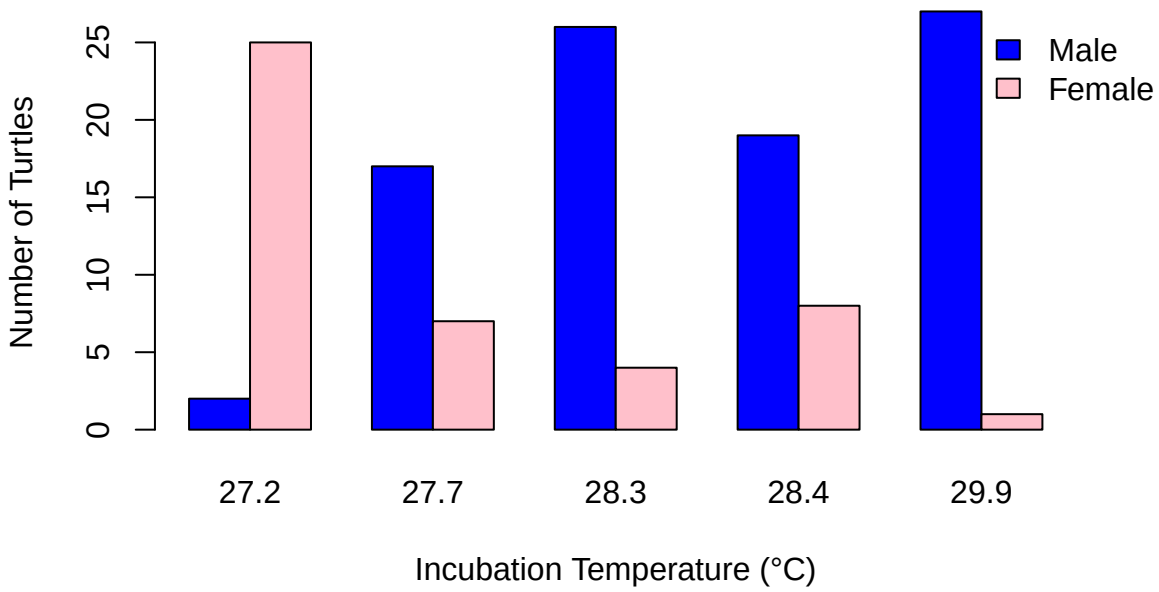
Environmental Statistics

Daniel Moses (5139055), Heiner Ochse (5741119) and Daniel Abanto (5706583)

Exercise 1

Subexercise 1.1

```
library(faraway)
data("turtle")
attach(turtle)


agg_data <- aggregate(cbind(turtle$male, female) ~ temp, data = turtle, sum)

# Create a bar plot
par(mar = c(5, 4, 4, 2))

barplot(
  t(as.matrix(agg_data[, -1])),
  beside = TRUE,
  col = c("blue", "pink"),
  names.arg = agg_data$temp,
  xlab = "Incubation Temperature (°C)",
  ylab = "Number of Turtles",
  main = "Sex Determination in Turtles",
  legend.text = c("Male", "Female"),
  args.legend = list(x ='topright', inset=c(-0.1,0), bty = "n")
)
```



Subexercise 1.2

```
# Models
modmale <- glm(cbind(male, female) ~ temp, data = turtle, family = "binomial")
summary(modmale)
```

```
Call:
glm(formula = cbind(male, female) ~ temp, family = "binomial",
    data = turtle)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0721  -1.0292  -0.2714   0.8087   2.5550

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -61.3183    12.0224  -5.100 3.39e-07 ***
temp          2.2110     0.4309   5.132 2.87e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.508  on 14  degrees of freedom
Residual deviance: 24.942  on 13  degrees of freedom
AIC: 53.836

Number of Fisher Scoring iterations: 5
```

```r
modfem <- glm(cbind(female, male) ~ temp, data = turtle, family = "binomial")
summary(modfem)
```

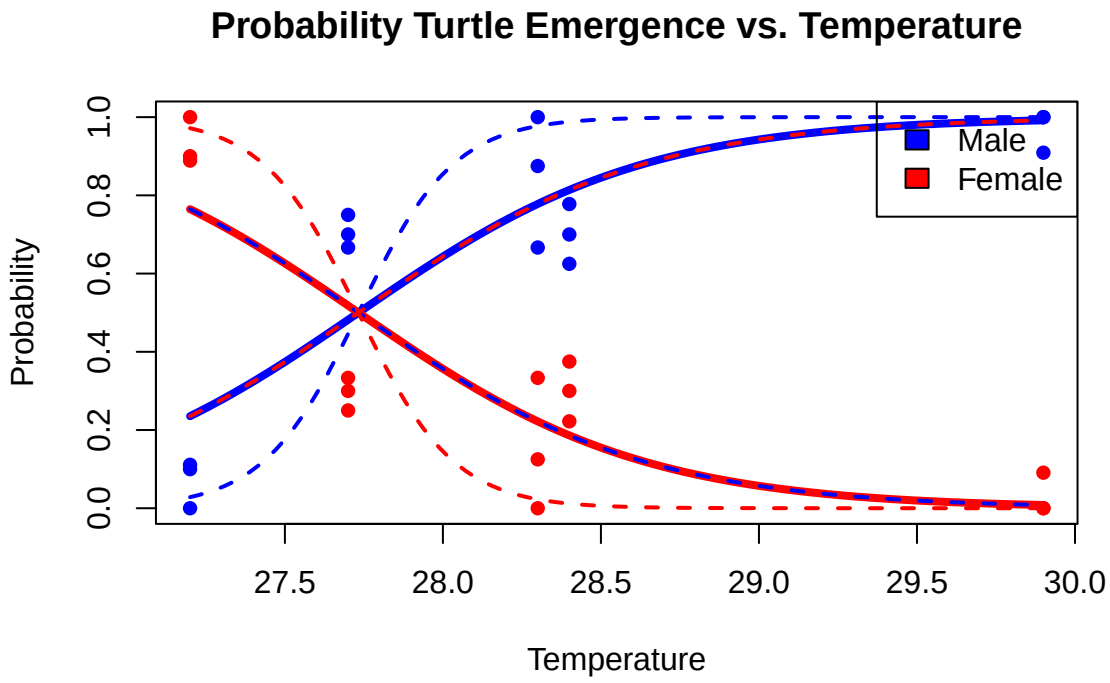
```
Call:
glm(formula = cbind(female, male) ~ temp, family = "binomial",
    data = turtle)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5550  -0.8087   0.2714   1.0292   2.0721

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  61.3183    12.0224   5.100 3.39e-07 ***
temp         -2.2110     0.4309  -5.132 2.87e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.508  on 14  degrees of freedom
Residual deviance: 24.942  on 13  degrees of freedom
AIC: 53.836

Number of Fisher Scoring iterations: 5
```

For the male model, the intercept is negative and the effect of temperature on the male gender is positive, i.e. with increasing temperature, the probability of male turtles hatching increases. For the female model it is the other way around, the intercept is positive and the effect of temperature on the female gender negative, so with increasing temperature, the probability of female turtles hatching decreases. In both cases, the effect of temperature was highly significant.

Subexercise 1.3

```r
new_data <- data.frame(temp = seq(min(turtle$temp), max(turtle$temp), length.out =
 ↪  100))
predicted_probsmale <- predict(modmale, newdata = new_data, se.fit= T)
predicted_probsfemale <- predict(modfem, newdata = new_data, se.fit=T)

# Plot the predicted probabilities
par(mar = c(5, 4, 4, 2))

plot(male / (female + male) ~ temp, pch = 16, col = "blue",
    xlab = "Temperature", ylab = "Probability",
    main = "Probability Turtle Emergence vs. Temperature")
points(female / (female + male) ~ temp, pch = 16, col = "red")
```

```
# Males
lines(new_data$temp, plogis(predicted_probsmale$fit), type = "l", lwd=4, col =
 ↪  "blue" )
# Females
lines(new_data$temp, plogis(predicted_probsfemale$fit), type = "l",lwd=4, col =
 ↪  "red")

# 95% confidence
lines(new_data$temp, plogis(predicted_probsmale$fit + 2 *
 ↪  predicted_probsmale$fit),
      type = "l", lty=2, lwd=2, col = "blue" )
lines(new_data$temp, plogis(predicted_probsmale$fit - 2 *
 ↪  predicted_probsmale$fit),
      type = "l", lty=2, lwd=2, col = "blue" )


lines(new_data$temp, plogis(predicted_probsfemale$fit + 2*
 ↪  predicted_probsfemale$fit),
      type = "l", lty= 2,lwd=2, col = "red")
lines(new_data$temp, plogis(predicted_probsfemale$fit - 2*
 ↪  predicted_probsfemale$fit),
      type = "l", lty= 2, lwd=2, col = "red")

legend("topright", legend = c("Male", "Female"), fill = c("blue", "red"))
```



Subexercise 1.4

```
xprob <- new_data$temp
yprob <- plogis(predicted_probsfemale$fit)

# Define the y-value for which you want to estimate the x-value
desired_y <- 0.5

# Use the `approx` function to estimate the x-value
estimated_x <- approx(yprob, xprob, xout = desired_y)$y

cat("Estimated x-value for y =", desired_y, "is x =", estimated_x, "\n")
```

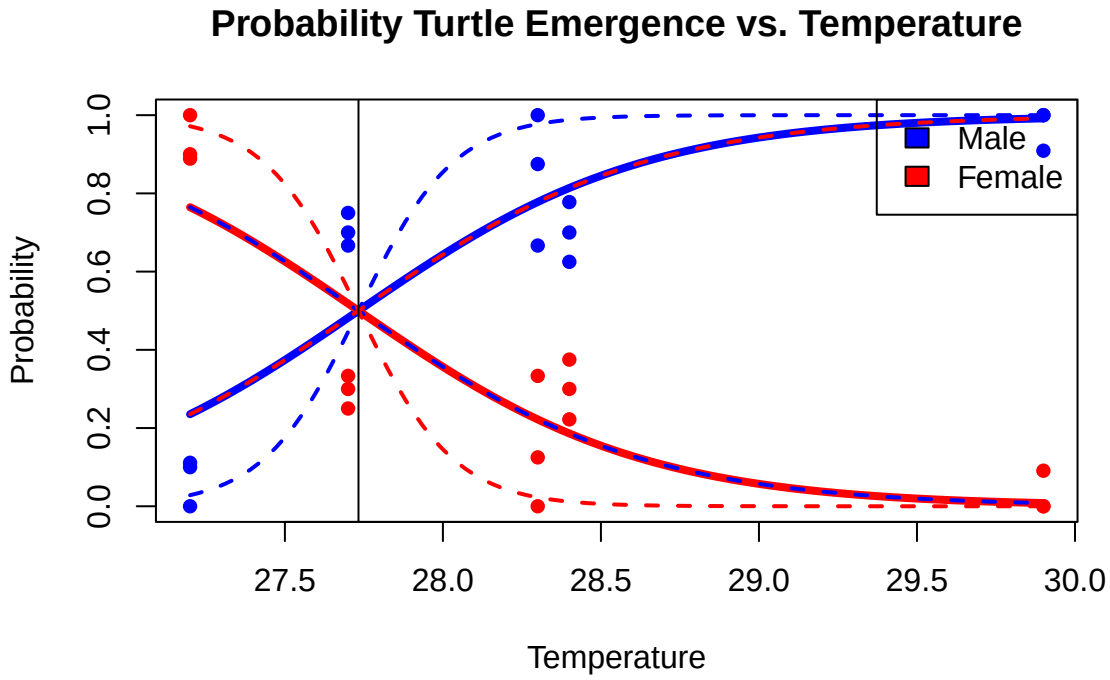
```
Estimated x-value for y = 0.5 is x = 27.7329
```

```
attach(turtle)
```

```
The following objects are masked from turtle (pos = 3):

    female, male, temp
```

```
#Plot with line at 50/50 point
par(mar = c(5, 4, 4, 2))

plot(male / (female + male) ~ temp, pch = 16, col = "blue",
    xlab = "Temperature", ylab = "Probability",
    main = "Probability Turtle Emergence vs. Temperature")
points(female / (female + male) ~ temp, pch = 16, col = "red")

# Males
lines(new_data$temp, plogis(predicted_probsmale$fit), type = "l", lwd=4, col =
 ↪  "blue" )
# Females
lines(new_data$temp, plogis(predicted_probsfemale$fit), type = "l",lwd=4, col =
 ↪  "red")

# 95% confidence
lines(new_data$temp, plogis(predicted_probsmale$fit + 2 *
 ↪  predicted_probsmale$fit),
      type = "l", lty=2, lwd=2, col = "blue" )
lines(new_data$temp, plogis(predicted_probsmale$fit - 2 *
 ↪  predicted_probsmale$fit),
      type = "l", lty=2, lwd=2, col = "blue" )
```

```r
lines(new_data$temp, plogis(predicted_probsfemale$fit + 2*
 ↪  predicted_probsfemale$fit),
      type = "l", lty= 2,lwd=2, col = "red")
lines(new_data$temp, plogis(predicted_probsfemale$fit - 2*
 ↪  predicted_probsfemale$fit),
      type = "l", lty= 2, lwd=2, col = "red")

legend("topright", legend = c("Male", "Female"), fill = c("blue", "red"))
abline(v=27.7329)
```



Instead of trying to find the intersect between the two lines, we tried to determine the temperature at which 50% of the hatched turtles are female, which would automatically be the 50/50 point between male and female.

Subexercise 1.5

The 50/50 point is very far to the left, i.e. the temperature is very low. This shows that over the course of the whole experiment, the probability of receiving a male turtle is higher than that of receiving a female turtle, because the regression line for male is above that for female for most of the time. At a temperature of 27,75 °C there is exactly a 50/50 chance that an egg will yield a female turtle or male turtle. The 50/50 chance of hatching can be seen at the intersect of the predicted male and female hatching curves. Regarding the plot it is remarkable to say that the 95% confidence interval of male hatching is the same as the regression line of female hatching after the intersection at 27.75 °C and viceversa.

Exercise 2

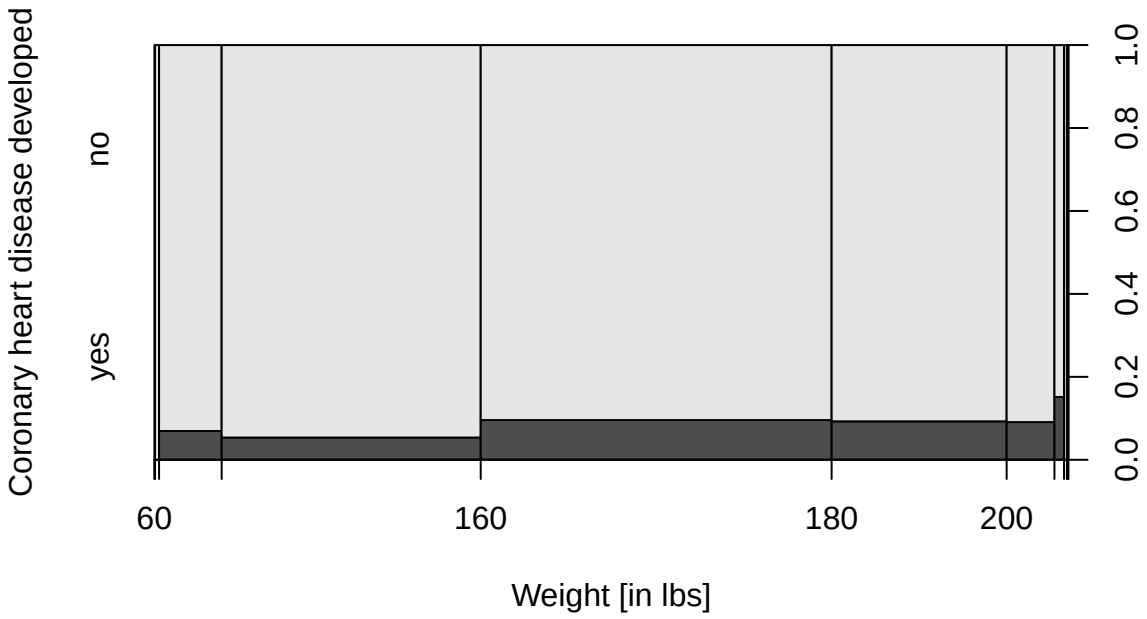
```
library(faraway)
data("wcgs")
attach(wcgs)
```

Subexercise 2.1

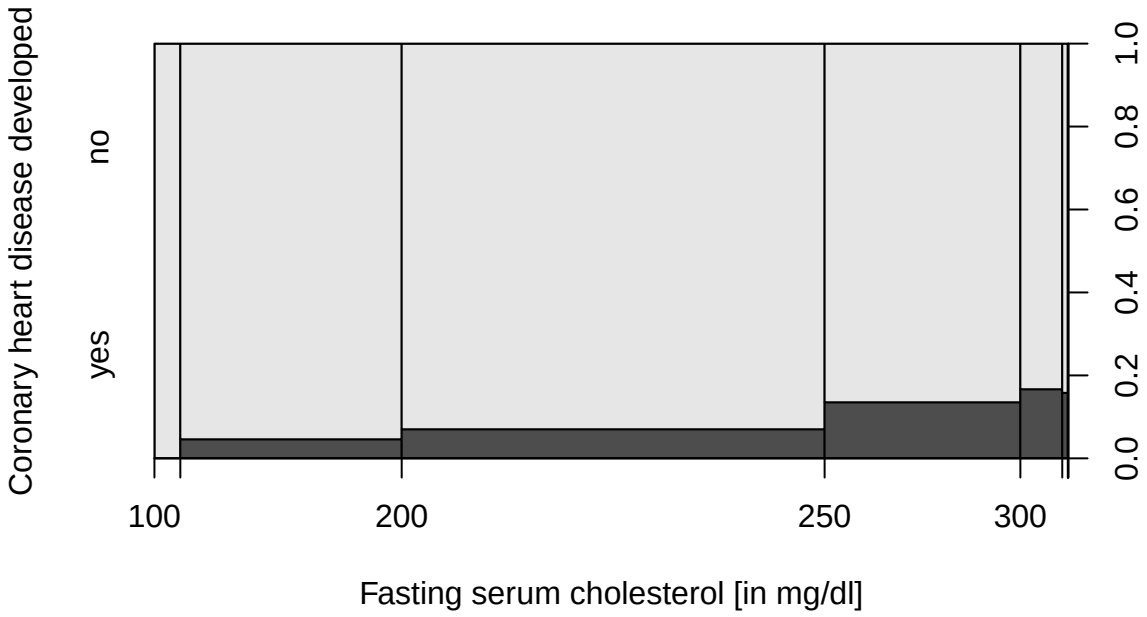
```
attach(wcgs)
```

```
The following objects are masked from wcgs (pos = 3):

    age, arcus, behave, chd, chol, cigs, dbp, dibep, height, sdp,
    timechd, typechd, weight
```

```
plot(chd ~ weight, xlab="Weight [in lbs]",
     ylab="Coronary heart disease developed")
```




```
plot(chd~chol, xlab="Fasting serum cholesterol [in mg/dl]",
     ylab= "Coronary heart disease developed")
```

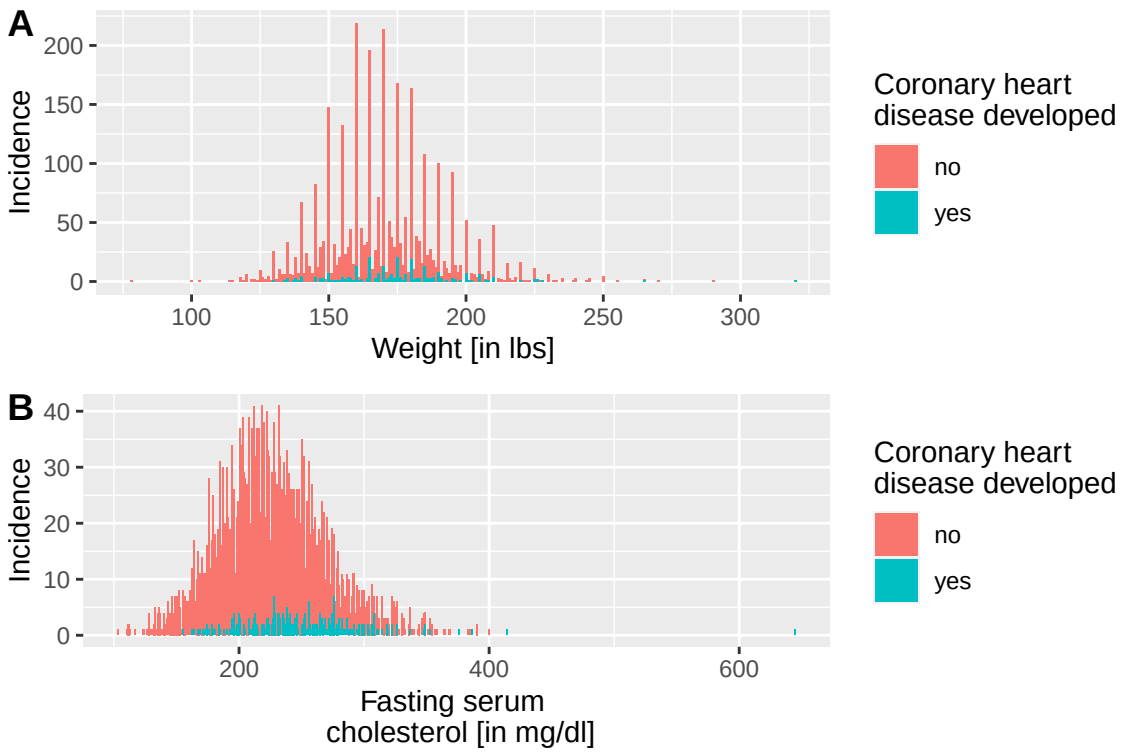


Incidence

```
library(ggplot2)
library(ggpubr)

g1 <- ggplot(wcgs, aes(x = weight, fill = chd)) +
  geom_bar(position = "stack") + labs(x="Weight [in lbs]",y="Incidence", fill =
                                        "Coronary heart \ndisease developed")


g2 <- ggplot(wcgs, aes(x = chol, fill = chd)) +
  geom_bar(position = "stack") + labs(x="Fasting serum \n cholesterol [in
   ↪  mg/dl]",y="Incidence", fill =
                                        "Coronary heart \ndisease developed")


z <- ggarrange(g1, g2,
          labels = c("A", "B"),
          ncol = 1, nrow = 2)
```

```
Warning: Removed 12 rows containing non-finite values (`stat_count()`).
```

```
z
```



Subexercise 2.2

```
attach(wcgs)
```

```
The following objects are masked from wcgs (pos = 5):

    age, arcus, behave, chd, chol, cigs, dbp, dibep, height, sdp,
    timechd, typechd, weight

The following objects are masked from wcgs (pos = 6):

    age, arcus, behave, chd, chol, cigs, dbp, dibep, height, sdp,
    timechd, typechd, weight
```

```
modchd <- glm(chd~weight + chol + weight*chol, family="binomial")

summary(modchd)
```

```
Call:
glm(formula = chd ~ weight + chol + weight * chol, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0996  -0.4499  -0.3644  -0.2886   2.7814
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.321e+01  2.580e+00  -5.121 3.03e-07 ***
weight       4.504e-02  1.449e-02   3.109 0.001876 **
chol         3.726e-02  1.038e-02   3.589 0.000333 ***
weight:chol -1.419e-04  5.864e-05  -2.420 0.015520 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1779.2  on 3141  degrees of freedom
Residual deviance: 1684.7  on 3138  degrees of freedom
  (12 observations deleted due to missingness)
AIC: 1692.7

Number of Fisher Scoring iterations: 5
```

Weight and cholesterol separately have a "positive" effect, i.e. with increasing weight and increasing cholesterol levels, more people will suffer from a coronary heart disease. Both effects are significant. The effect of the interaction between weight and cholesterol levels on coronary heart diseases is estimated to be negative. This suggests that the interaction is associated with a decrease in coronary heart diseases, because the effect of e.g. weight on coronary heart diseases may be different depending on the level of cholesterol and this interaction has a negative impact on coronary heart diseases as compared to what we would expect based on the main effects alone. As in this case a negative effect indicates a decrease in coronary heart diseases it could be because if the cholesterol levels are lower, the effect of weight isn't as strong anymore and viceversa.

```
modchd$deviance/modchd$df.residual
```

```
[1] 0.5368713
```

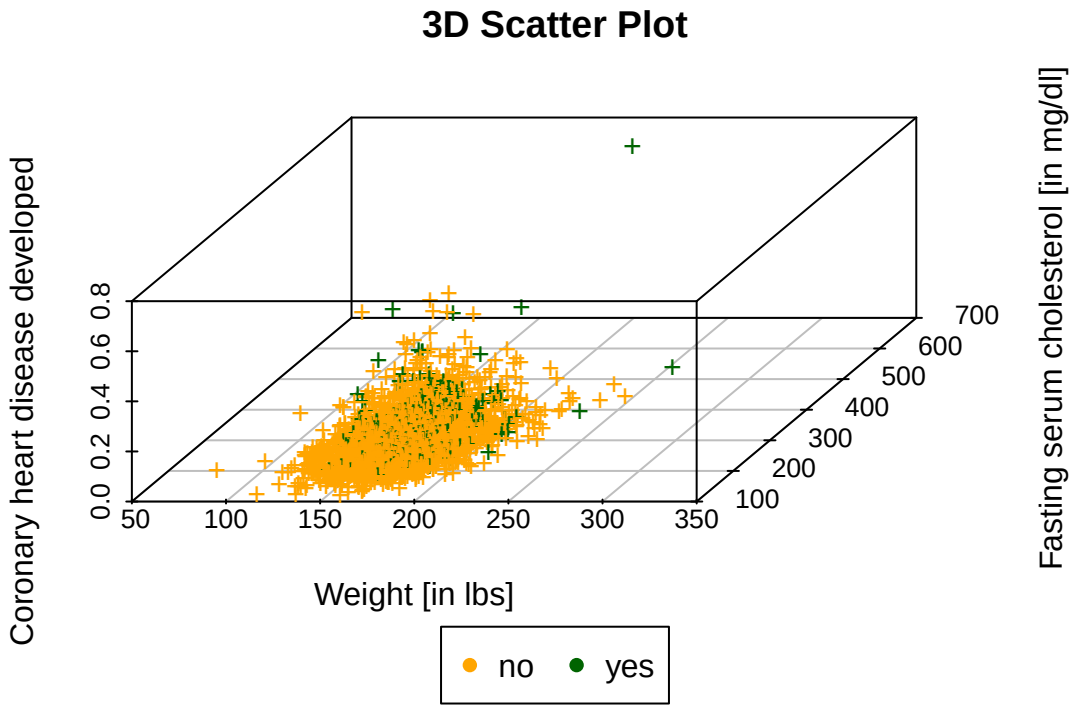
The dispersion parameter is lower than 1, indicating underdispersion. This means that the observed variance in the response variable (i.e. coronary heart diseases) is less than expected based on the model. This suggests that the model is overly conservative in its estimates of variability.

Subexercise 2.3

```
library(faraway)
data("wcgs")
library(scatterplot3d)
#install.packages("scatterplot3d")

sum(is.na(wcgs$chol))
```

```
[1] 12
```

```r
wcgs <- wcgs[!is.na(wcgs$chol),]

# Generate a linear model
mod <- glm(chd~weight + chol + weight*chol, family="binomial", data = wcgs)


# Get predicted values of mpg from wt and disp
wcgs$pred_chd_1 <- predict(mod, type = "response")

#2 Farben:
colors <- c("orange", "darkgreen")
colors <- colors[as.numeric(wcgs$chd)]

scatterplot3d(wcgs$weight, wcgs$chol, wcgs$pred_chd_1, type="p", xlab="Weight [in
 ↪  lbs]",
              ylab="Fasting serum cholesterol [in mg/dl]",
              angle=55,
              zlab= "Coronary heart disease developed",
              pch = "+", color = colors,
              main = "3D Scatter Plot")

legend("bottom", legend = levels(wcgs$chd),
       col =  c("orange", "darkgreen"), pch = 16,
       inset = -0.45, xpd = TRUE, horiz = TRUE)
```



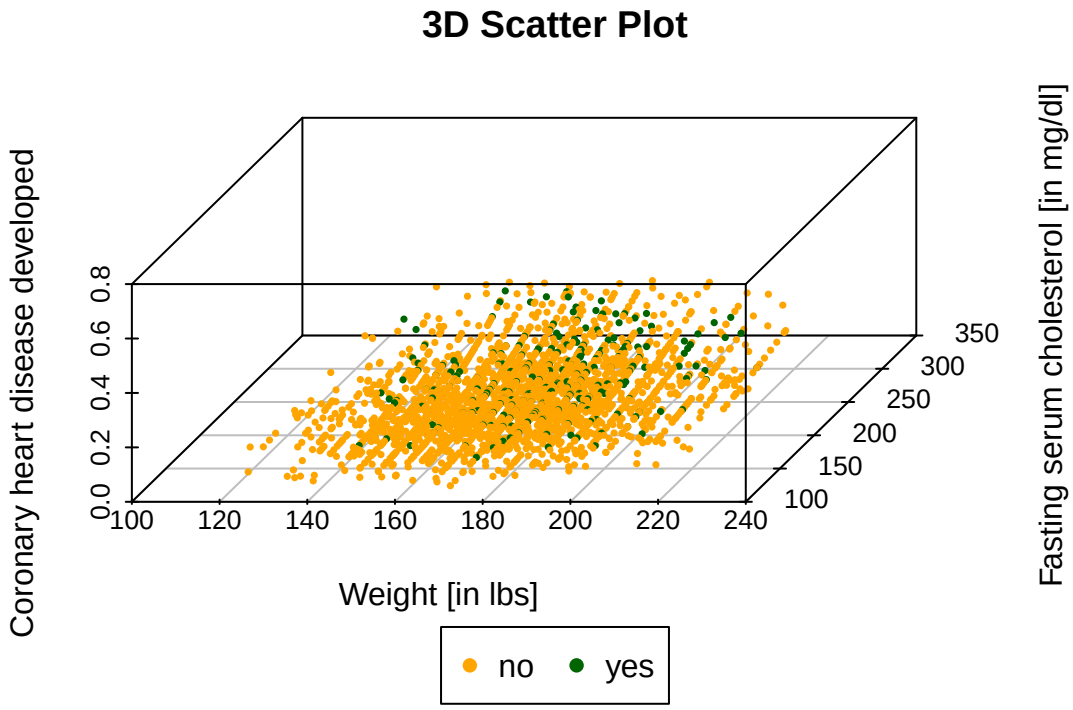

```r
# Fit a logistic regression model to predict CHD using weight and chol as
 ↪  predictors
```

```r
library(faraway)
data("wcgs")
library(scatterplot3d)


sum(is.na(wcgs$chol))
```

```
[1] 12
```

```r
wcgs <- wcgs[!is.na(wcgs$chol),]

# Generate a linear model
mod <- glm(chd~weight + chol + weight*chol, family="binomial", data = wcgs)


# Get predicted values of mpg from wt and disp
wcgs$pred_chd_2 <- predict(mod, type = "response")


# Define a function to remove outliers
remove_outliers <- function(x, na.rm = TRUE) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm)
  H <- 1.5 * IQR(x, na.rm = na.rm)
  x[x < (qnt[1] - H)] <- NA
  x[x > (qnt[2] + H)] <- NA
  x
}

colors <- c("orange", "darkgreen")
colors <- colors[as.numeric(wcgs$chd)]

shapes = c(17, 18)
shapes <- shapes[as.numeric(wcgs$chd)]

# Remove outliers for weight and chol
wcgs$weight <- remove_outliers(wcgs$weight)
wcgs$chol <- remove_outliers(wcgs$chol)


scatterplot3d(wcgs$weight, wcgs$chol, wcgs$pred_chd_2,
              pch = 16, angle=55, type = "p", cex.symbols = 0.5,
              xlab="Weight [in lbs]",
              ylab="Fasting serum cholesterol [in mg/dl]",
              zlab= "Coronary heart disease developed",
              color = colors,
```

```
              box = TRUE,
              main = "3D Scatter Plot")


legend("bottom", legend = levels(wcgs$chd),
       col =  c("orange", "darkgreen"), pch = 16,
       inset = -0.45, xpd = TRUE, horiz = TRUE)
```



Subexercise 2.4

All coefficients are statistically significant. Weight as predictor has a positive coefficient in the glm. That means that with increasing weight, the probability of suffering from a coronary heart disease increases (holding other variables constant). Similarly, weight also has a positive coefficient in the glm, meaning that higher cholesterol levels are associated with a higher probability of suffering from a coronary heart disease (holding other variables constant). As the coefficient for the interaction term is negative, the effect of weight on the probability of suffering from a coronary heart disease is weaker as cholesterol levels increase. In other words, the effect of weight is moderated by the level of cholesterol. As cholesterol levels increase, the effect of weight on the probability of coronary heart disease becomes weaker. This can be seen in the 3D-plots.