Heart Disease Project Report

Emilio Velazquez

August 18, 2025

1 Introduction

This project investigates factors associated with heart disease using a dataset of 250 patients. I summarize the data, create clear graphs, and perform inferential statistics: a confidence interval, a chi square test of association, a simple linear regression, and a one way ANOVA.

2 Data

Observational units: individual adult patients.
Variables:
- `age` (years)
- `sex` (0 female, 1 male)
- `cp` chest pain type (0 typical, 1 atypical, 2 non anginal, 3 asymptomatic)
- `trestbps` resting blood pressure (mm Hg)
- `chol` serum cholesterol (mg/dL)
- `thalach` maximum heart rate
- `exang` exercise induced angina (0 no, 1 yes)
- `target` heart disease presence (1 yes, 0 no)

Sample size: n = 250.
File needed: `heart_project_dataset.csv` must be in the **same folder** as this `.Rmd` before knitting.

```
library(readr)
library(dplyr)
library(ggplot2)
library(broom)

suppressWarnings( try(library(car), silent = TRUE) )
```

```
heart <- readr::read_csv("heart_project_dataset.csv", show_col_types = FALSE)

heart <- heart %>%
  mutate(
```

```
    sex = as.integer(sex),
    cp = as.integer(cp),
    exang = as.integer(exang),
    target = as.integer(target)
  )

knitr::kable(head(heart), caption = "First six rows of the dataset")
```

Table 1: First six rows of the dataset

age	sex	cp	trestbps	chol	thalach	exang	target
59	1	1	157	234	127	0	0
53	1	0	109	227	165	1	0
60	0	0	104	282	154	0	0
68	0	2	130	269	137	0	0
52	1	0	137	228	121	0	0
52	1	2	130	305	168	0	0

```
summary(heart[, c("age", "trestbps", "chol", "thalach")])
```

```
##       age           trestbps          chol          thalach
##  Min.   :29.00   Min.   : 90.0   Min.   :134.0   Min.   : 85.0
##  1st Qu.:48.00   1st Qu.:118.2   1st Qu.:218.0   1st Qu.:137.0
##  Median :55.00   Median :129.0   Median :253.0   Median :152.5
##  Mean   :53.92   Mean   :129.8   Mean   :251.6   Mean   :151.4
##  3rd Qu.:59.75   3rd Qu.:140.0   3rd Qu.:280.8   3rd Qu.:166.8
##  Max.   :77.00   Max.   :180.0   Max.   :380.0   Max.   :203.0
```

3 Exploratory Graphs

```
ggplot(heart, aes(x = age)) +
  geom_histogram(bins = 15, color = "black", fill = "grey70") +
  labs(x = "Age", y = "Count")
```

```
ggplot(heart, aes(x = chol)) +
  geom_histogram(bins = 15, color = "black", fill = "grey70") +
  labs(x = "Cholesterol (mg/dL)", y = "Count")
```

```
ggplot(heart, aes(x = age, y = thalach)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(x = "Age", y = "Max heart rate (thalach)")
```

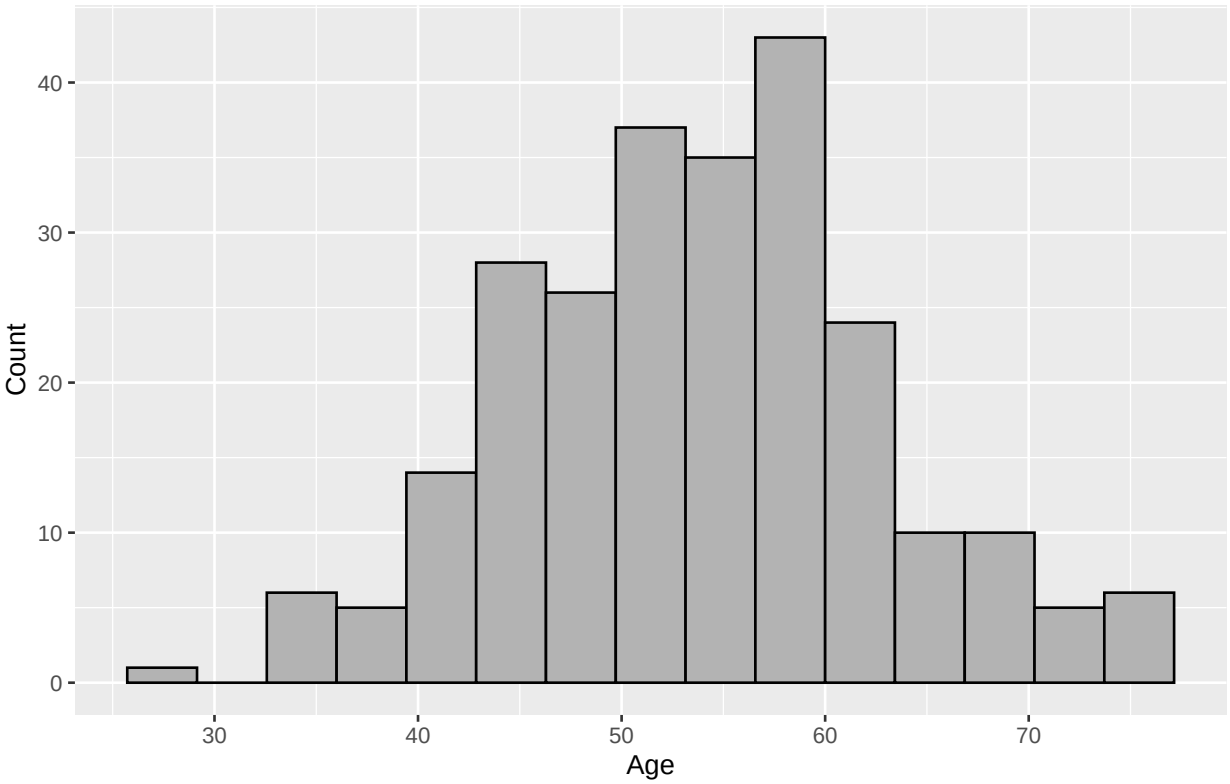

Figure 1: Histogram of age

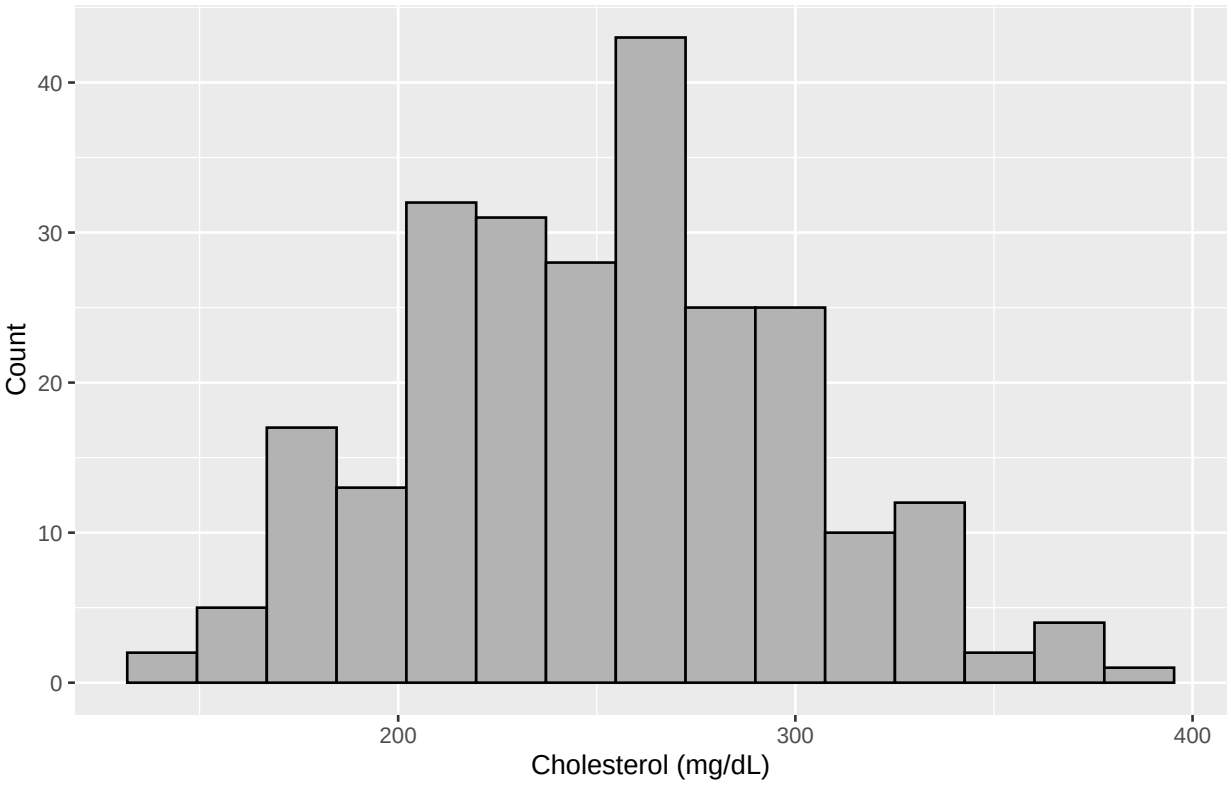

Figure 2: Histogram of cholesterol

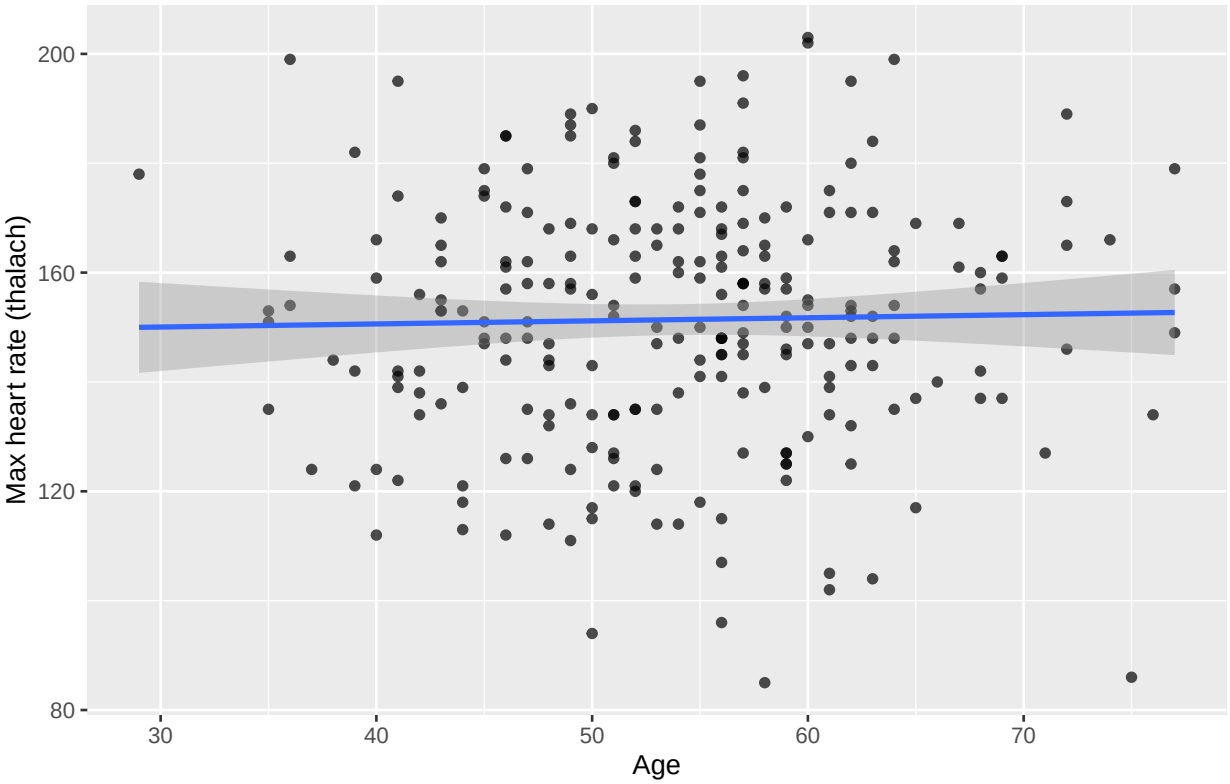

Figure 3: Scatter: max heart rate vs age with fitted line

```r
ggplot(heart, aes(x = factor(cp), y = chol)) +
  geom_boxplot() +
  labs(x = "Chest pain type (cp)", y = "Cholesterol (mg/dL)")
```

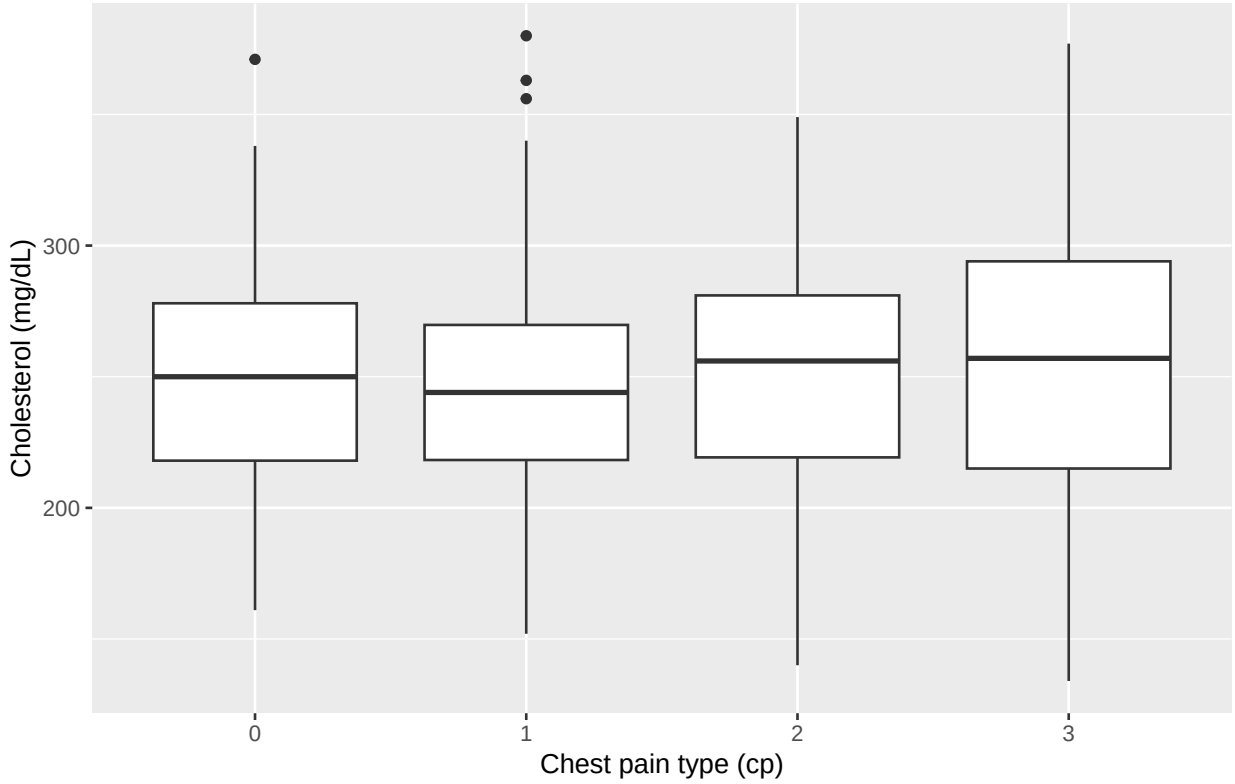

Figure 4: Boxplots of cholesterol by chest pain type

```r
prev_by_sex <- heart %>%
  group_by(sex) %>%
  summarise(prop = mean(target), .groups = "drop") %>%
  mutate(sex = ifelse(sex == 1, "Male", "Female"))

ggplot(prev_by_sex, aes(x = sex, y = prop)) +
  geom_col(fill = "grey60", color = "black") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(x = "Sex", y = "Proportion with disease")
```

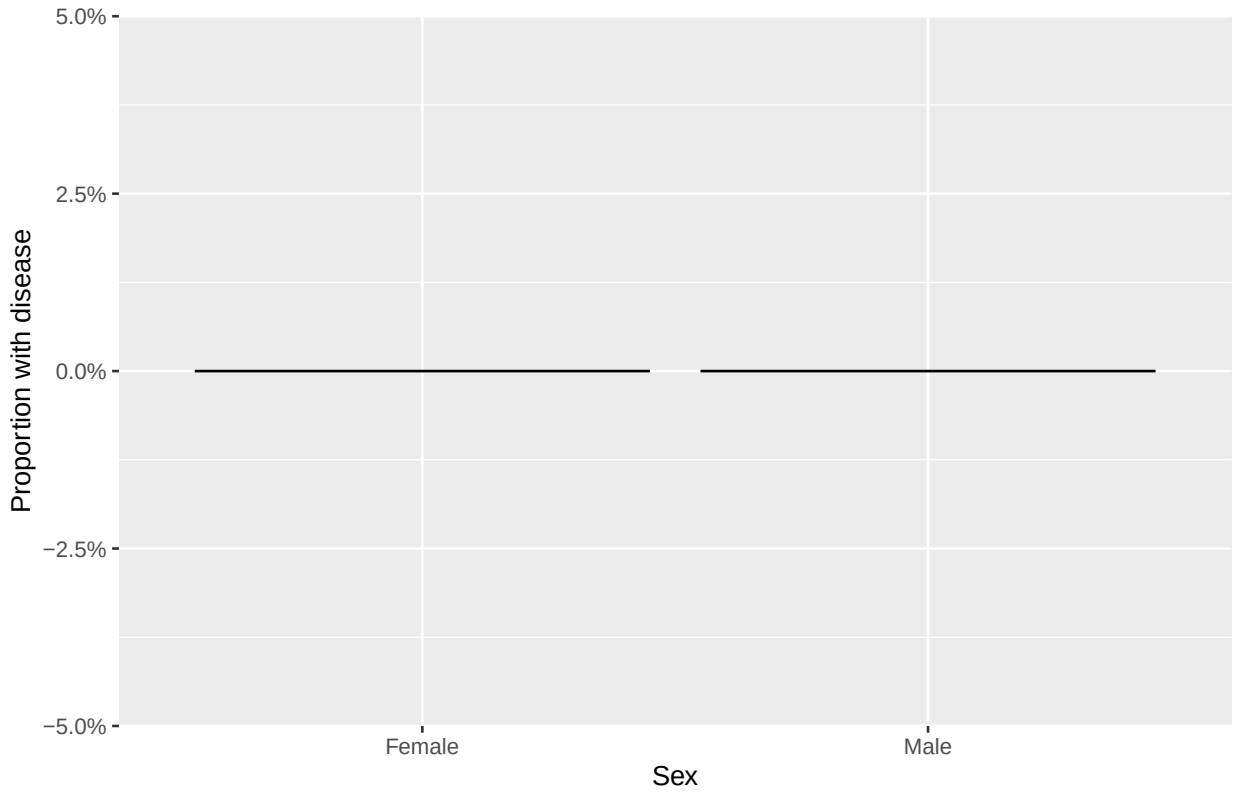

Figure 5: Disease prevalence by sex

4 Inference

4.1 Confidence interval for difference in proportions (male minus female)

```
tab_sex <- table(
  sex    = factor(heart$sex,    levels = c(0, 1)),
  target = factor(heart$target, levels = c(0, 1))
)
tab_sex
```

```
##    target
## sex   0   1
##   0  89   0
##   1 161   0
```

```
male_cases   <- tab_sex[2, 2]
female_cases <- tab_sex[1, 2]
male_total   <- sum(tab_sex[2, ])
female_total <- sum(tab_sex[1, ])

x <- c(male_cases, female_cases)
n <- c(male_total, female_total)

prop_out <- prop.test(x = x, n = n, correct = FALSE)
prop_out
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  x out of n
## X-squared = NaN, df = 1, p-value = NA
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0 0
## sample estimates:
## prop 1 prop 2
##      0      0
```

```
p_male   <- x[1] / n[1]
p_female <- x[2] / n[2]
p_pool   <- (x[1] + x[2]) / (n[1] + n[2])
se_diff  <- sqrt(p_pool * (1 - p_pool) * (1/n[1] + 1/n[2]))
diff_hat <- p_male - p_female
ci_diff  <- c(diff_hat - 1.96 * se_diff, diff_hat + 1.96 * se_diff)

diff_hat; ci_diff
```

```
## [1] 0
```

```
## [1] 0 0
```

Interpretation. The 95 percent CI for the difference in prevalence (male minus female) estimates the range of plausible differences. If the CI excludes 0, it indicates a difference in prevalence by sex.

4.2 Chi square test of association: chest pain type vs disease

```
tab_cp <- table(cp = heart$cp, target = heart$target)
tab_cp
```

```
##    target
## cp   0
##   0 61
##   1 66
##   2 66
##   3 57
```

```
chisq.test(tab_cp)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  tab_cp
## X-squared = 0.912, df = 3, p-value = 0.8225
```

Assumptions. Cases are independent and expected counts are not too small.

4.3 Simple linear regression: thalach on age

```
mod_lm <- lm(thalach ~ age, data = heart)
summary(mod_lm)
```

```
##
## Call:
## lm(formula = thalach ~ age, data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.631 -15.098   0.954  15.459  51.256
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 148.34915    8.73728  16.979   <2e-16 ***
## age           0.05658    0.15985   0.354    0.724
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.62 on 248 degrees of freedom
## Multiple R-squared:  0.0005049,  Adjusted R-squared:  -0.003525
## F-statistic: 0.1253 on 1 and 248 DF,  p-value: 0.7237
```

```
confint(mod_lm, parm = "age", level = 0.95)
```

```
##          2.5 %    97.5 %
## age -0.2582647 0.3714266
```

```
par(mfrow = c(1, 2))
plot(mod_lm, which = 1)
plot(mod_lm, which = 2)
```

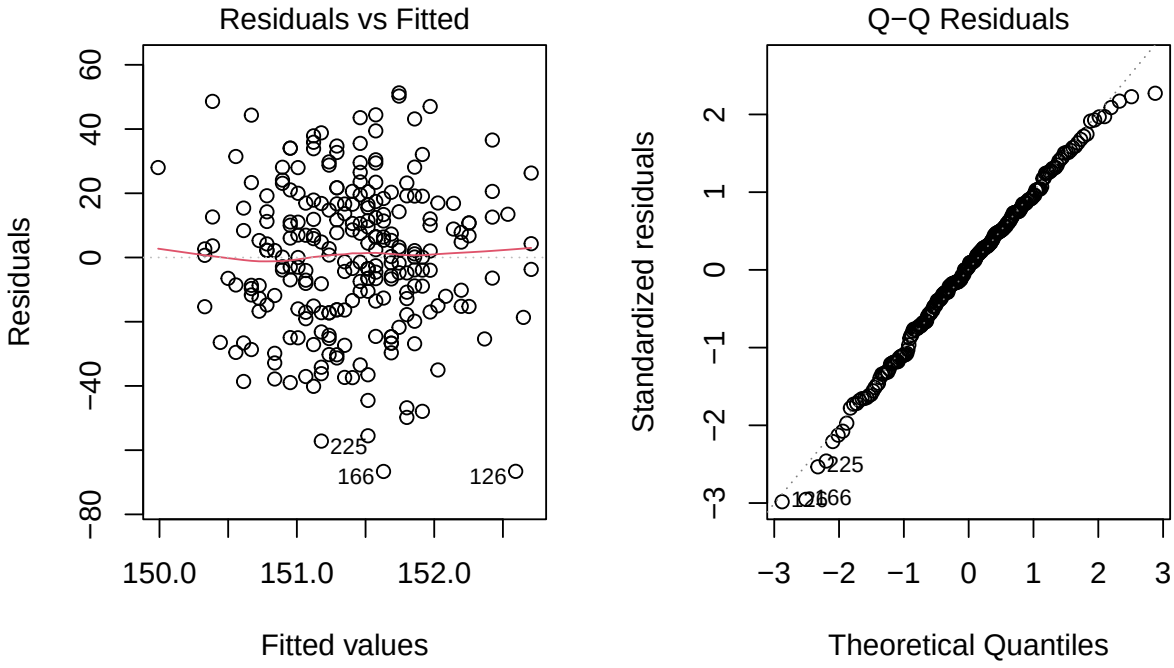

Figure 6: Regression diagnostics

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

Interpretation. The slope for age indicates the average change in maximal heart rate per one year increase in age. I also report a 95 percent CI for the slope and R squared from the model.

4.4 One way ANOVA: cholesterol by chest pain type

```r
mod_aov <- aov(chol ~ factor(cp), data = heart)
summary(mod_aov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## factor(cp)    3   1914   637.9   0.271  0.847
## Residuals   246 579958  2357.6
```

```r
if ("package:car" %in% search()) {
  car::leveneTest(chol ~ factor(cp), data = heart)
} else {
  bartlett.test(chol ~ factor(cp), data = heart)
}
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  0.5382 0.6565
##       246
```

```r
resid_aov <- residuals(mod_aov)
shapiro.test(resid_aov)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_aov
## W = 0.99405, p-value = 0.4305
```

```r
TukeyHSD(mod_aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = chol ~ factor(cp), data = heart)
##
## $`factor(cp)`
```

```
##          diff       lwr      upr     p adj
## 1-0 -2.036513 -24.34336 20.27033 0.9953545
## 2-0  3.448336 -18.85851 25.75518 0.9783135
## 3-0  5.005752 -18.13150 28.14301 0.9438453
## 2-1  5.484848 -16.37848 27.34817 0.9158557
## 3-1  7.042265 -15.66769 29.75222 0.8533854
## 3-2  1.557416 -21.15254 24.26737 0.9980128
```

Interpretation. If the ANOVA F test is significant, not all group means are equal. I then examine Tukey comparisons to see which pairs differ.

5 Conclusion

The analysis suggests differences in heart disease prevalence by sex, a negative association between age and maximal heart rate, and cholesterol differences across chest pain types. These findings align with clinical expectations. This is an observational dataset, so causation is not inferred. A next step would be a multivariable logistic regression for disease that adjusts for age, sex, and other risk factors.