BUDA 530 Assignment 2

Collin Edwards

2025-02-10

You are a data scientist working for a consulting firm specializing in safety and public health. Your current projects involve analyzing data from the Titanic passenger manifest and the Framingham Heart Study. Your boss has assigned you to respond to specific client questions and prepare deliverables for upcoming meetings.

Data

The `titanic` dataset is from the `datasets` library. This dataset contains information on the passengers of the Titanic. Use `help("Titanic")` for more information.

The `framingham` dataset come from Kaggle (.CSV included on eCampus). You can find more information on it at Logistic regression To predict heart disease. It is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD).The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

Scenarios and Tasks:

Scenario 1: Titanic Data Analysis Client Email Request: A cruise company executive has sent the following email:
"Dear Analyst,
We are in the process of designing a new luxury cruise liner. To enhance passenger safety, we would like to understand the factors that influenced survival during the Titanic disaster. Specifically, we are curious about how age, gender, and ticket class affected survival rates. Could you provide a summary of insights and any recommendations?

Best regards,
Executive Team"

Your Task: 1. Analyze the Titanic dataset to identify how survival rates varied by age, gender, and ticket class. - Use a binomial logistic regression model to predict the survival of a passenger using all the other variables as predictors.

1.0 Analyzing the dataset

```
data("Titanic")
str(Titanic) # to look at the structure of the data
```

```
##  'table' num [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
##  - attr(*, "dimnames")=List of 4
##   ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
##   ..$ Sex     : chr [1:2] "Male" "Female"
##   ..$ Age     : chr [1:2] "Child" "Adult"
##   ..$ Survived: chr [1:2] "No" "Yes"
```

```r
# Converting Titanic table to a Data Frame
titanic_df <- as.data.frame(Titanic)
# using View() to open it in a spreadsheet-like view. It's a contingency table so it's not a standard d
# View(your_dataframe)  # Commented out since View() doesn't work in RMarkdown
head(titanic_df)  # Use this instead for previewing data
```

```
##   Class    Sex   Age Survived Freq
## 1   1st   Male Child       No    0
## 2   2nd   Male Child       No    0
## 3   3rd   Male Child       No   35
## 4  Crew   Male Child       No    0
## 5   1st Female Child       No    0
## 6   2nd Female Child       No    0
```

```r
summary(titanic_df)
```

```
##   Class       Sex         Age     Survived      Freq
##  1st :8   Male  :16   Child:16   No :16   Min.   :  0.00
##  2nd :8   Female:16   Adult:16   Yes:16   1st Qu.:  0.75
##  3rd :8                                  Median : 13.50
##  Crew:8                                  Mean   : 68.78
##                                          3rd Qu.: 77.00
##                                          Max.   :670.00
```

1.1 Binomial logistic regression model

```r
# Binomial logistic regression model
# fitting the logistic regression model
titanic_mod1 <- glm(Survived ~ Class + Sex + Age, #using the glm instead of the regular lm function sin
                    data = titanic_df,
                    weights = Freq,
                    family = binomial)

# Display the model summary
summary(titanic_mod1)
```

```
##
## Call:
## glm(formula = Survived ~ Class + Sex + Age, family = binomial,
##     data = titanic_df, weights = Freq)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.6853     0.2730   2.510   0.0121 *
## Class2nd     -1.0181     0.1960  -5.194 2.05e-07 ***
## Class3rd     -1.7778     0.1716 -10.362  < 2e-16 ***
## ClassCrew    -0.8577     0.1573  -5.451 5.00e-08 ***
## SexFemale     2.4201     0.1404  17.236  < 2e-16 ***
## AgeAdult     -1.0615     0.2440  -4.350 1.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2769.5  on 23  degrees of freedom
## Residual deviance: 2210.1  on 18  degrees of freedom
## AIC: 2222.1
##
## Number of Fisher Scoring iterations: 5
```

We now fit a binomial logistic regression model. In this model, the response variable is Survived and the predictors are Class, Sex, and Age. Because the data is aggregated, we include the Freq column as weights. The model is specified using the glm() function with the binomial family.

1.2 Stepwise model selection I will now use the step() function (with both forward and backward selection) to identify a more parsimonious model based on AIC. We don't want a very complex model so the AIC will penalize models by giving them a higher score. We only use AIC when we're looking at sub models of the same family. This process will consider potential interactions and may exclude variables that do not contribute significantly—even if there are reasons to keep them for subject-matter purposes.

```
# Stepwise model selection
titanic_mod2 <- step(titanic_mod1, direction = "both", trace = TRUE)
```

```
## Start:  AIC=2222.06
## Survived ~ Class + Sex + Age
##
##         Df Deviance    AIC
## <none>       2210.1 2222.1
## - Age    1   2228.9 2238.9
## - Class  3   2329.1 2335.1
## - Sex    1   2563.0 2573.0
```

```
# Display the final model summary
summary(titanic_mod2)
```

```
##
## Call:
## glm(formula = Survived ~ Class + Sex + Age, family = binomial,
##     data = titanic_df, weights = Freq)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.6853     0.2730   2.510   0.0121 *
## Class2nd     -1.0181     0.1960  -5.194 2.05e-07 ***
## Class3rd     -1.7778     0.1716 -10.362  < 2e-16 ***
## ClassCrew    -0.8577     0.1573  -5.451 5.00e-08 ***
## SexFemale     2.4201     0.1404  17.236  < 2e-16 ***
## AgeAdult     -1.0615     0.2440  -4.350 1.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2769.5  on 23  degrees of freedom
## Residual deviance: 2210.1  on 18  degrees of freedom
```

```
## AIC: 2222.1
##
## Number of Fisher Scoring iterations: 5
```

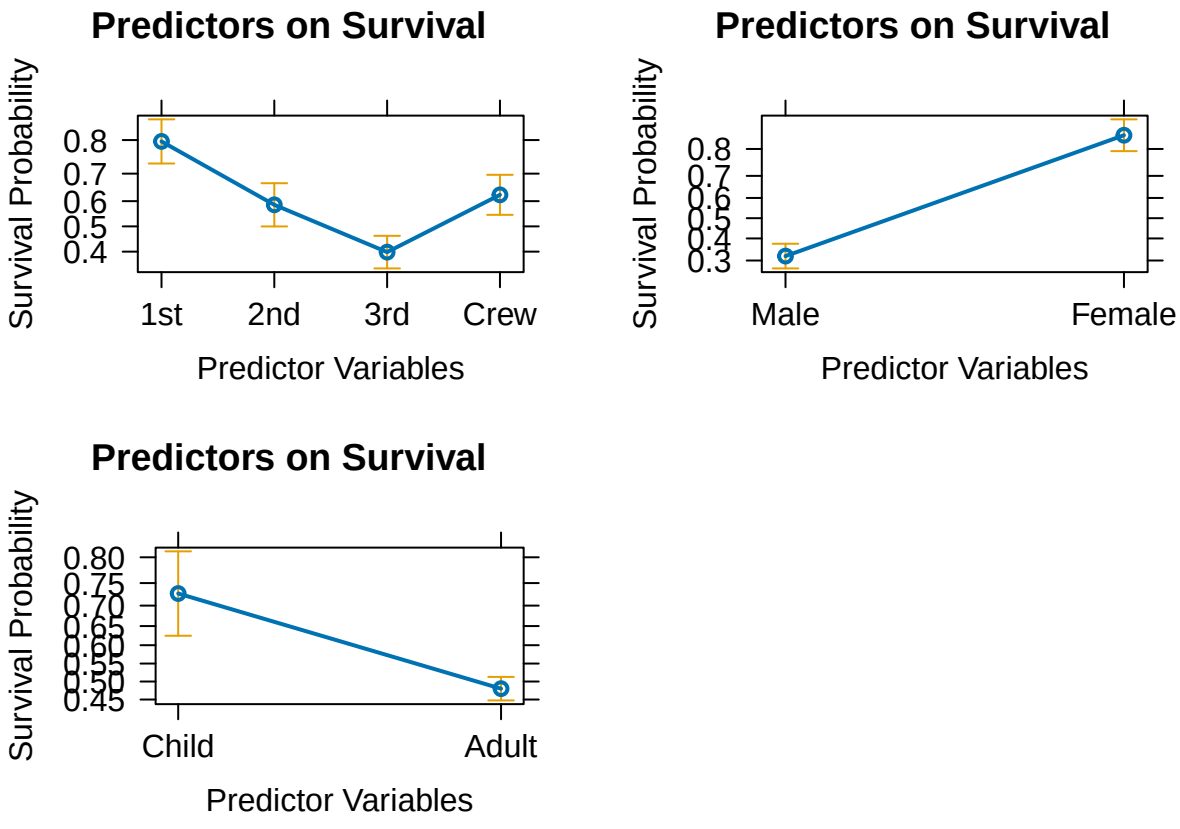
1.3 Visualizations

```
# Load the effects package
library(effects)
```

```
## Loading required package: carData
```

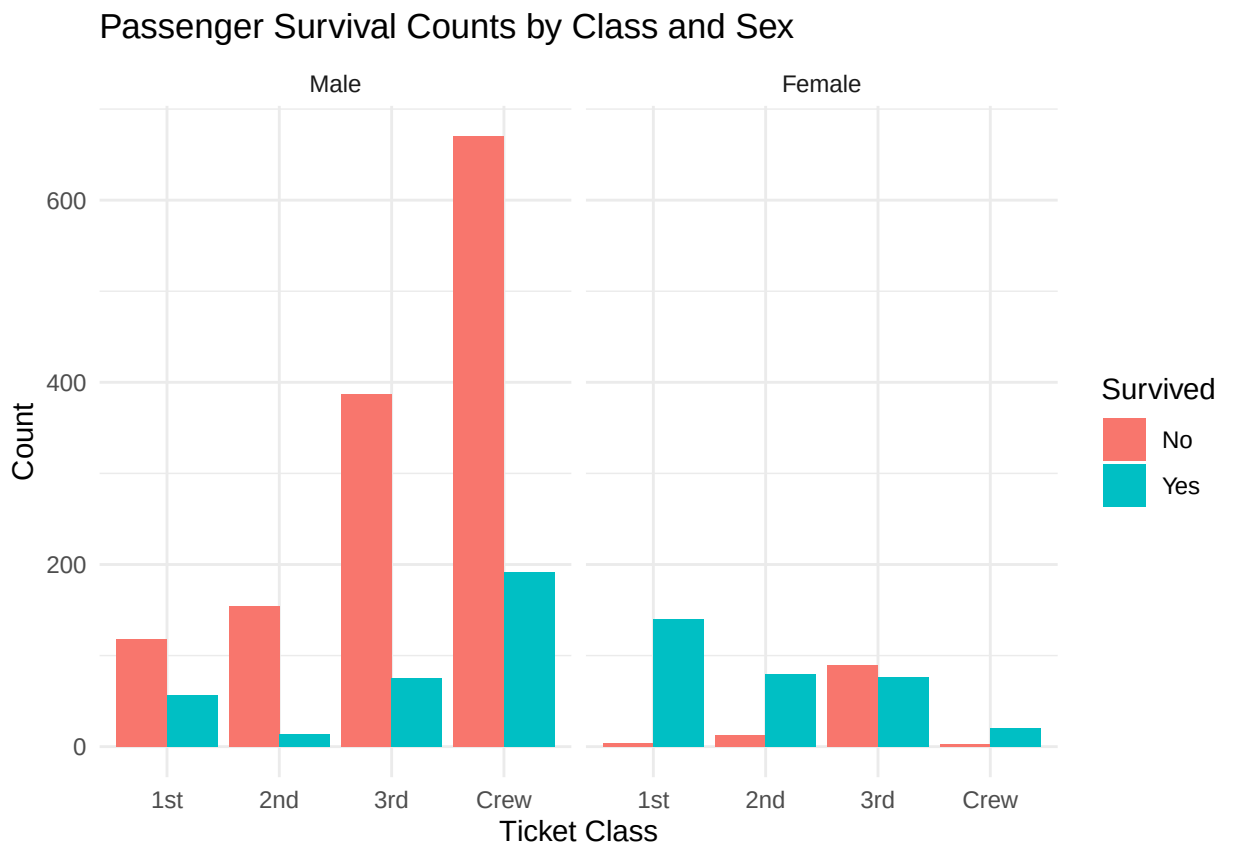
```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
# Plot the effects of the predictors in the final model
plot(allEffects(titanic_mod2),
     main = "Predictors on Survival",
     xlab = "Predictor Variables",
     ylab = "Survival Probability",
     fig.height = 6,
     fig.width = 8)
```



1.3.5 Alternative visualization

```r
library(ggplot2)
ggplot(titanic_df, aes(x = Class, y = Freq, fill = Survived)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~Sex) +
  labs(title = "Passenger Survival Counts by Class and Sex",
       x = "Ticket Class",
       y = "Count") +
  theme_minimal()
```



Report Email **Subject:** Insights on Passenger Survival Factors – Titanic Analysis

Dear Executive Team,

Thank you for reaching out regarding passenger safety insights based on the Titanic disaster. Our analysis of the dataset revealed key factors influencing survival rates: **Gender:** Female passengers had significantly higher survival odds compared to males. **Ticket Class:** Passengers in first class had the highest survival rates, followed by second and third class, with third-class passengers facing the lowest odds of survival. **Age:** Children had better survival chances than adults, though the effect was less pronounced than gender and class. **Recommendations for Enhancing Passenger Safety:** Prioritized Emergency Protocols: Implement clear, structured emergency procedures ensuring equitable access to lifeboats across all passenger classes. Training & Drills: Regular crew training and passenger safety drills can improve response times during emergencies. Enhanced Safety Infrastructure: Consider improved life-saving equipment allocation and strategically placed emergency exits to ensure accessibility for all passengers, particularly those in economy-class accommodations. These insights can help inform safety measures for your new cruise liner, ensuring improved preparedness and survival outcomes in emergency situations.

Best regards,

Collin Edwards

Scenario 2: Framingham Heart Study

Town Hall Prep: Your boss is preparing for a town hall meeting with public health officials. She has asked you to prepare an answer to the following question:
"What are the most critical demographic and lifestyle factors contributing to cardiovascular disease risk in our region? Are there specific interventions we should focus on?"

Your Task: 1. Analyze the Framingham dataset to identify key risk factors (e.g., age, cholesterol levels, smoking habits). - Complete an analysis of this data using a binary logistic regression model to predict the 10 year risk of coronary heart disease using all the other variables as predictors. - Use the `step` function to select the best model. - Note: you will need to deal with "NAs". - Note: you may need to consider the interactions between certain predictor variables (`currentSmoker` and `cigsPerDay`). - Consider whether some variables should be excluded for reasons other than model fit. 2. Create clear visualizations to illustrate your findings. - Use the `effects` package to plot the effects of the predictors in the final model. - Remember to adjust the `fig.height` and `fig.width` chunk options so that the plots look nice). 3. Prepare a detailed written response, including: - A summary of the most significant risk factors. - Recommendations for targeted interventions, with specific examples based on your analysis.

2.1 Preparing and analyzing the dataset

```
# Load the Framingham dataset
framingham <- read.csv("~/Downloads/framingham.csv")
# Check the structure of the dataset
str(framingham)
```

```
## 'data.frame':    4238 obs. of  16 variables:
##  $ male           : int  1 0 1 0 0 0 0 0 1 1 ...
##  $ age            : int  39 46 48 61 46 43 63 45 52 43 ...
##  $ education      : int  4 2 1 3 3 2 1 2 1 1 ...
##  $ currentSmoker  : int  0 0 1 1 1 0 0 1 0 1 ...
##  $ cigsPerDay     : int  0 0 20 30 23 0 0 20 0 30 ...
##  $ BPMeds         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentHyp   : int  0 0 0 1 0 1 0 0 1 1 ...
##  $ diabetes       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ totChol        : int  195 250 245 225 285 228 205 313 260 225 ...
##  $ sysBP          : num  106 121 128 150 130 ...
##  $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
##  $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
##  $ heartRate      : int  80 95 75 65 85 77 60 79 76 93 ...
##  $ glucose        : int  77 76 70 103 85 99 85 78 79 88 ...
##  $ TenYearCHD     : int  0 0 0 1 0 0 1 0 0 0 ...
```

```
# Check for missing values
sum(is.na(framingham))
```

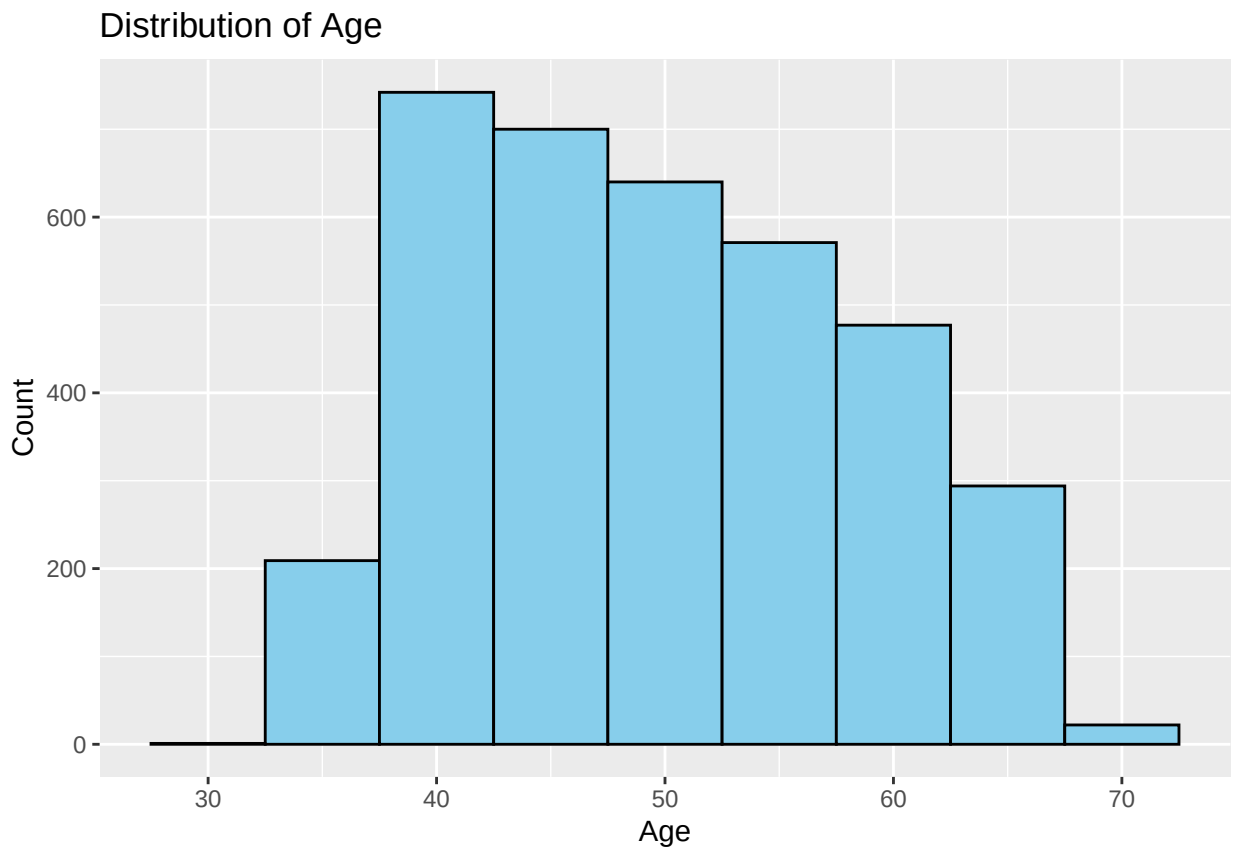
```
## [1] 645
```

```r
# Remove rows with missing values
framingham <- na.omit(framingham)
summary(framingham)
```

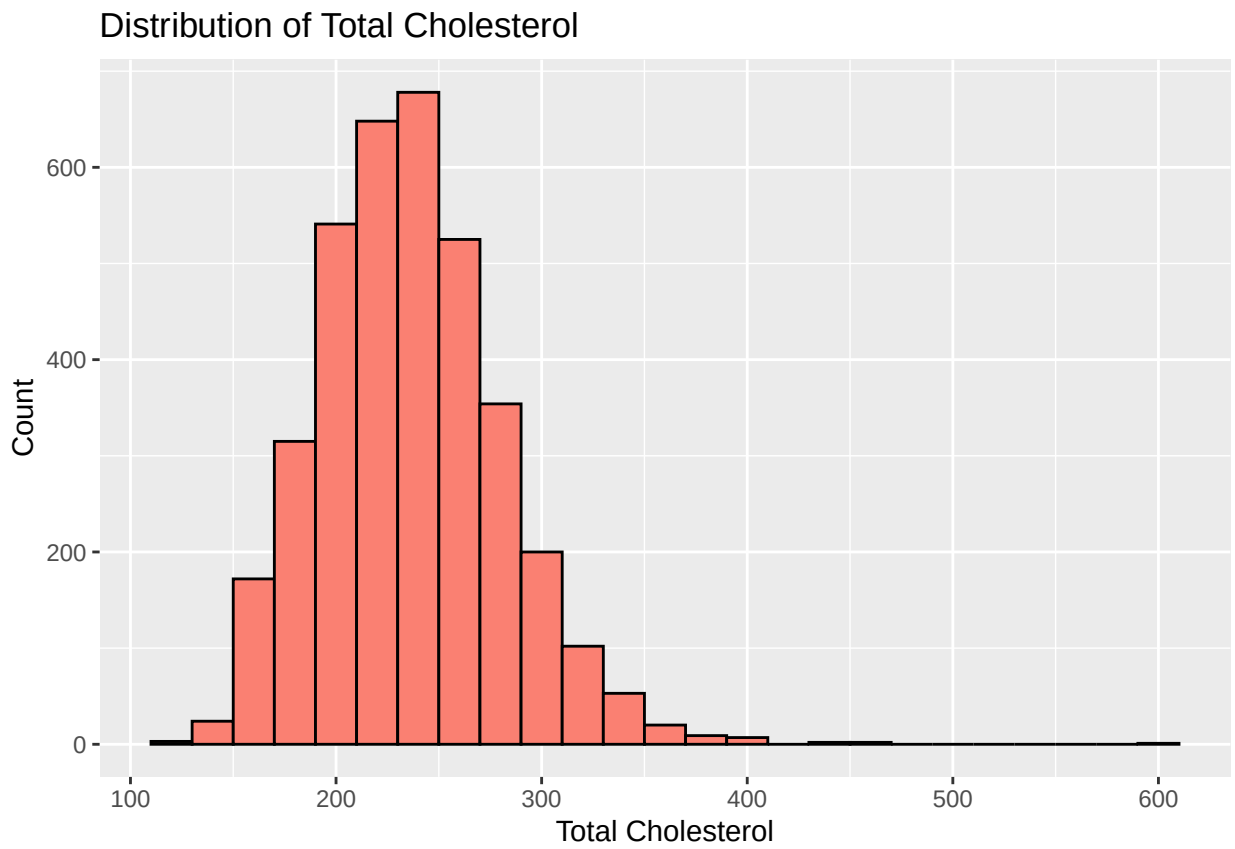
```
##       male             age          education    currentSmoker
##  Min.   :0.0000   Min.   :32.00   Min.   :1.00   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.00   1st Qu.:0.0000
##  Median :0.0000   Median :49.00   Median :2.00   Median :0.0000
##  Mean   :0.4437   Mean   :49.56   Mean   :1.98   Mean   :0.4891
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.00   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :70.00   Max.   :4.00   Max.   :1.0000
##    cigsPerDay        BPMeds        prevalentStroke     prevalentHyp
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000
##  Mean   : 9.022   Mean   :0.03036   Mean   :0.005744   Mean   :0.3115
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
##     diabetes          totChol          sysBP           diaBP
##  Min.   :0.00000   Min.   :113.0   Min.   : 83.5   Min.   : 48.00
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.00
##  Median :0.00000   Median :234.0   Median :128.0   Median : 82.00
##  Mean   :0.02708   Mean   :236.9   Mean   :132.4   Mean   : 82.91
##  3rd Qu.:0.00000   3rd Qu.:263.2   3rd Qu.:144.0   3rd Qu.: 90.00
##  Max.   :1.00000   Max.   :600.0   Max.   :295.0   Max.   :142.50
##       BMI          heartRate         glucose         TenYearCHD
##  Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.0000
##  1st Qu.:23.08   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.0000
##  Median :25.38   Median : 75.00   Median : 78.00   Median :0.0000
##  Mean   :25.78   Mean   : 75.73   Mean   : 81.86   Mean   :0.1524
##  3rd Qu.:28.04   3rd Qu.: 82.00   3rd Qu.: 87.00   3rd Qu.:0.0000
##  Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.0000
```

2.1.1 Exploring the dataset

```r
# plotting the distribution of age and total cholesterol
ggplot(framingham, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Age", x = "Age", y = "Count")
```




```
ggplot(framingham, aes(x = totChol)) +
  geom_histogram(binwidth = 20, fill = "salmon", color = "black") +
  labs(title = "Distribution of Total Cholesterol", x = "Total Cholesterol", y = "Count")
```



2.1.2 Binomial logistic regression model

```
# Binary logistic regression model
# Fitting the logistic regression model
framingham_mod <- glm(TenYearCHD ~ . + currentSmoker:cigsPerDay,
                  data = framingham,
                  family = binomial)

# Summarize the full model
summary(framingham_mod)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ . + currentSmoker:cigsPerDay, family = binomial,
##     data = framingham)
##
## Coefficients: (1 not defined because of singularities)
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -8.322206   0.715476 -11.632  < 2e-16 ***
## male                     0.555098   0.109046   5.091 3.57e-07 ***
## age                      0.063453   0.006680   9.499  < 2e-16 ***
## education               -0.047497   0.049390  -0.962  0.33621
## currentSmoker            0.070875   0.156749   0.452  0.65115
## cigsPerDay               0.017929   0.006238   2.874  0.00405 **
## BPMeds                   0.162255   0.234309   0.692  0.48863
## prevalentStroke          0.693502   0.489532   1.417  0.15658
```

```
## prevalentHyp              0.234638   0.138037   1.700  0.08917 .
## diabetes                  0.039461   0.315483   0.125  0.90046
## totChol                   0.002324   0.001127   2.062  0.03920 *
## sysBP                     0.015398   0.003808   4.043 5.27e-05 ***
## diaBP                    -0.004132   0.006438  -0.642  0.52096
## BMI                       0.006603   0.012758   0.518  0.60476
## heartRate                -0.003250   0.004211  -0.772  0.44030
## glucose                   0.007124   0.002234   3.189  0.00143 **
## currentSmoker:cigsPerDay        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3120.5  on 3655  degrees of freedom
## Residual deviance: 2754.2  on 3640  degrees of freedom
## AIC: 2786.2
##
## Number of Fisher Scoring iterations: 5
```

We now fit a full logistic regression model where the response variable is `TenYearCHD` and all other variables are used as predictors. In addition, we explicitly include an interaction term between `currentSmoker` and `cigsPerDay` to capture the combined effect of smoking status and smoking intensity.

2.1.3 Stepwise model selection

```r
# Stepwise model selection
framingham_mod_step <- step(framingham_mod, direction = "both", trace = TRUE)
```

```
## Start:  AIC=2786.2
## TenYearCHD ~ male + age + education + currentSmoker + cigsPerDay +
##     BPMeds + prevalentStroke + prevalentHyp + diabetes + totChol +
##     sysBP + diaBP + BMI + heartRate + glucose + currentSmoker:cigsPerDay
##
##
## Step:  AIC=2786.2
## TenYearCHD ~ male + age + education + currentSmoker + cigsPerDay +
##     BPMeds + prevalentStroke + prevalentHyp + diabetes + totChol +
##     sysBP + diaBP + BMI + heartRate + glucose
##
##                   Df Deviance    AIC
## - diabetes         1   2754.2 2784.2
## - currentSmoker    1   2754.4 2784.4
## - BMI              1   2754.5 2784.5
## - diaBP            1   2754.6 2784.6
## - BPMeds           1   2754.7 2784.7
## - heartRate        1   2754.8 2784.8
## - education        1   2755.1 2785.1
## - prevalentStroke  1   2756.1 2786.1
## <none>                 2754.2 2786.2
## - prevalentHyp     1   2757.1 2787.1
## - totChol          1   2758.4 2788.4
## - cigsPerDay       1   2762.4 2792.4
## - glucose          1   2764.7 2794.7
```

```
## - sysBP            1   2770.5 2800.5
## - male             1   2780.3 2810.3
## - age              1   2847.6 2877.6
##
## Step:  AIC=2784.21
## TenYearCHD ~ male + age + education + currentSmoker + cigsPerDay +
##     BPMeds + prevalentStroke + prevalentHyp + totChol + sysBP +
##     diaBP + BMI + heartRate + glucose
##
##                   Df Deviance    AIC
## - currentSmoker    1   2754.4 2782.4
## - BMI              1   2754.5 2782.5
## - diaBP            1   2754.6 2782.6
## - BPMeds           1   2754.7 2782.7
## - heartRate        1   2754.8 2782.8
## - education        1   2755.2 2783.2
## - prevalentStroke  1   2756.1 2784.1
## <none>                 2754.2 2784.2
## - prevalentHyp     1   2757.1 2785.1
## + diabetes         1   2754.2 2786.2
## - totChol          1   2758.4 2786.4
## - cigsPerDay       1   2762.4 2790.4
## - sysBP            1   2770.6 2798.6
## - glucose          1   2773.0 2801.0
## - male             1   2780.4 2808.4
## - age              1   2847.7 2875.7
##
## Step:  AIC=2782.42
## TenYearCHD ~ male + age + education + cigsPerDay + BPMeds + prevalentStroke +
##     prevalentHyp + totChol + sysBP + diaBP + BMI + heartRate +
##     glucose
##
##                   Df Deviance    AIC
## - BMI              1   2754.6 2780.6
## - diaBP            1   2754.8 2780.8
## - BPMeds           1   2754.9 2780.9
## - heartRate        1   2755.0 2781.0
## - education        1   2755.3 2781.3
## - prevalentStroke  1   2756.3 2782.3
## <none>                 2754.4 2782.4
## - prevalentHyp     1   2757.3 2783.3
## + currentSmoker    1   2754.2 2784.2
## + diabetes         1   2754.4 2784.4
## - totChol          1   2758.6 2784.6
## - sysBP            1   2770.8 2796.8
## - glucose          1   2773.2 2799.2
## - cigsPerDay       1   2776.3 2802.3
## - male             1   2780.5 2806.5
## - age              1   2847.8 2873.8
##
## Step:  AIC=2780.64
## TenYearCHD ~ male + age + education + cigsPerDay + BPMeds + prevalentStroke +
##     prevalentHyp + totChol + sysBP + diaBP + heartRate + glucose
##
```

```
##                   Df Deviance    AIC
## - diaBP            1   2755.0 2779.0
## - BPMeds           1   2755.1 2779.1
## - heartRate        1   2755.2 2779.2
## - education        1   2755.7 2779.7
## - prevalentStroke  1   2756.6 2780.6
## <none>                 2754.6 2780.6
## - prevalentHyp     1   2757.6 2781.6
## + BMI              1   2754.4 2782.4
## + currentSmoker    1   2754.5 2782.5
## + diabetes         1   2754.6 2782.6
## - totChol          1   2758.9 2782.9
## - sysBP            1   2771.0 2795.0
## - glucose          1   2773.7 2797.7
## - cigsPerDay       1   2776.3 2800.3
## - male             1   2780.9 2804.9
## - age              1   2847.9 2871.9
##
## Step:  AIC=2778.97
## TenYearCHD ~ male + age + education + cigsPerDay + BPMeds + prevalentStroke +
##     prevalentHyp + totChol + sysBP + heartRate + glucose
##
##                   Df Deviance    AIC
## - BPMeds           1   2755.5 2777.5
## - heartRate        1   2755.6 2777.6
## - education        1   2756.1 2778.1
## - prevalentStroke  1   2756.9 2778.9
## <none>                 2755.0 2779.0
## - prevalentHyp     1   2757.7 2779.7
## + diaBP            1   2754.6 2780.6
## + currentSmoker    1   2754.8 2780.8
## + BMI              1   2754.8 2780.8
## + diabetes         1   2754.9 2780.9
## - totChol          1   2759.2 2781.2
## - glucose          1   2774.4 2796.4
## - cigsPerDay       1   2776.9 2798.9
## - sysBP            1   2778.2 2800.2
## - male             1   2780.9 2802.9
## - age              1   2854.7 2876.7
##
## Step:  AIC=2777.49
## TenYearCHD ~ male + age + education + cigsPerDay + prevalentStroke +
##     prevalentHyp + totChol + sysBP + heartRate + glucose
##
##                   Df Deviance    AIC
## - heartRate        1   2756.2 2776.2
## - education        1   2756.6 2776.6
## <none>                 2755.5 2777.5
## - prevalentStroke  1   2757.6 2777.6
## - prevalentHyp     1   2758.5 2778.5
## + BPMeds           1   2755.0 2779.0
## + diaBP            1   2755.1 2779.1
## + currentSmoker    1   2755.3 2779.3
## + BMI              1   2755.3 2779.3
```

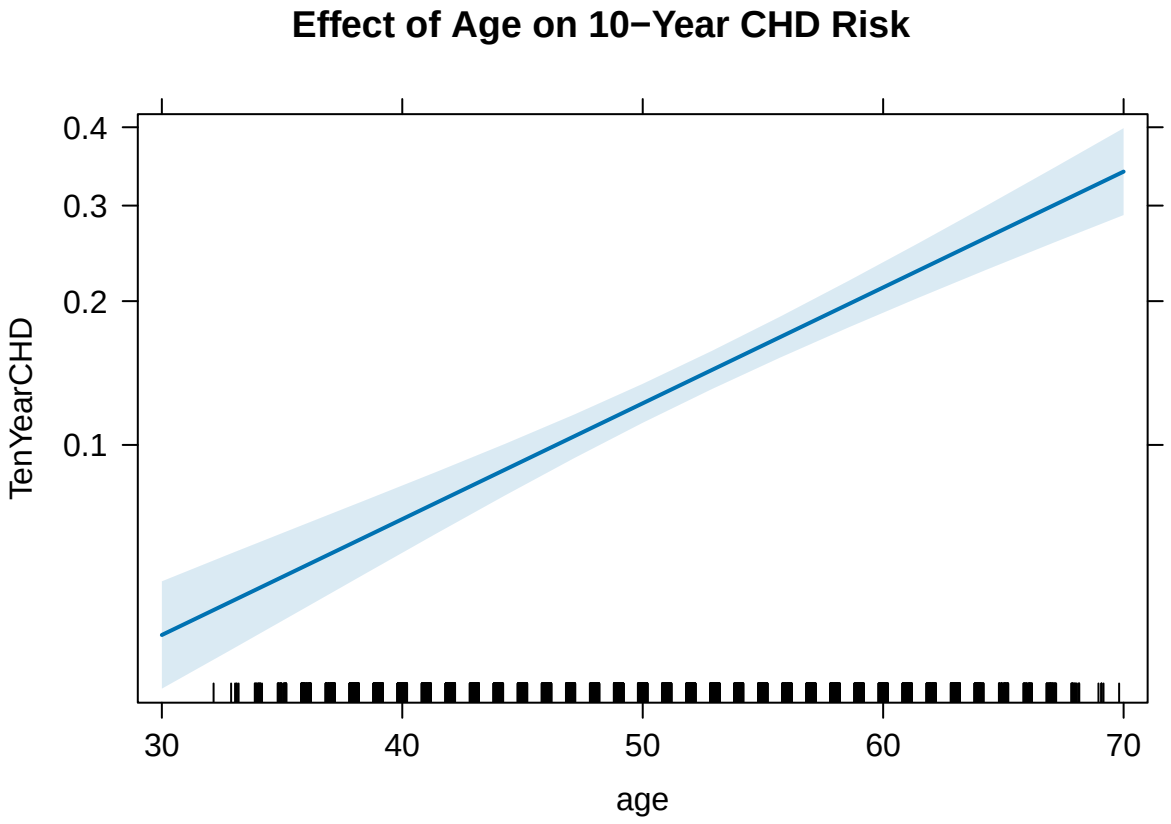
```
## + diabetes         1   2755.5 2779.5
## - totChol          1   2759.8 2779.8
## - glucose          1   2775.0 2795.0
## - cigsPerDay       1   2777.4 2797.4
## - sysBP            1   2780.1 2800.1
## - male             1   2781.2 2801.2
## - age              1   2855.5 2875.5
##
## Step:  AIC=2776.15
## TenYearCHD ~ male + age + education + cigsPerDay + prevalentStroke +
##     prevalentHyp + totChol + sysBP + glucose
##
##                    Df Deviance    AIC
## - education         1   2757.2 2775.2
## <none>                  2756.2 2776.2
## - prevalentStroke   1   2758.4 2776.4
## - prevalentHyp      1   2759.0 2777.0
## + heartRate         1   2755.5 2777.5
## + BPMeds            1   2755.6 2777.6
## + diaBP             1   2755.8 2777.8
## + currentSmoker     1   2756.0 2778.0
## + BMI               1   2756.0 2778.0
## + diabetes          1   2756.1 2778.1
## - totChol           1   2760.3 2778.3
## - glucose           1   2775.2 2793.2
## - cigsPerDay        1   2777.4 2795.4
## - sysBP             1   2780.2 2798.2
## - male              1   2783.2 2801.2
## - age               1   2858.2 2876.2
##
## Step:  AIC=2775.19
## TenYearCHD ~ male + age + cigsPerDay + prevalentStroke + prevalentHyp +
##     totChol + sysBP + glucose
##
##                    Df Deviance    AIC
## <none>                  2757.2 2775.2
## - prevalentStroke   1   2759.5 2775.5
## - prevalentHyp      1   2760.0 2776.0
## + education         1   2756.2 2776.2
## + heartRate         1   2756.6 2776.6
## + BPMeds            1   2756.7 2776.7
## + diaBP             1   2756.7 2776.7
## + BMI               1   2757.0 2777.0
## + currentSmoker     1   2757.0 2777.0
## + diabetes          1   2757.2 2777.2
## - totChol           1   2761.2 2777.2
## - glucose           1   2776.4 2792.4
## - cigsPerDay        1   2778.7 2794.7
## - sysBP             1   2782.0 2798.0
## - male              1   2784.1 2800.1
## - age               1   2863.9 2879.9
```

```r
# Summarize the final model
summary(framingham_mod_step)
```

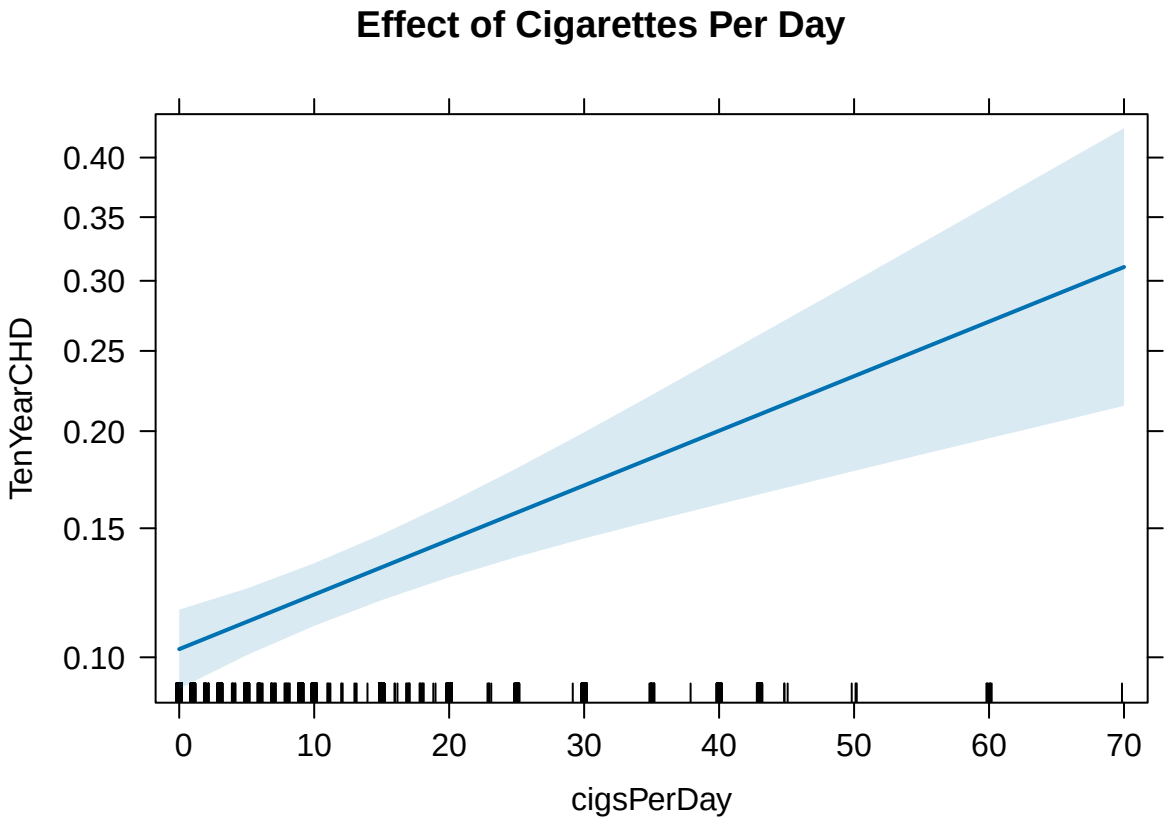
```
##
## Call:
## glm(formula = TenYearCHD ~ male + age + cigsPerDay + prevalentStroke +
##     prevalentHyp + totChol + sysBP + glucose, family = binomial,
##     data = framingham)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.739521   0.522563 -16.724  < 2e-16 ***
## male             0.553152   0.107037   5.168 2.37e-07 ***
## age              0.065337   0.006444  10.140  < 2e-16 ***
## cigsPerDay       0.019574   0.004182   4.681 2.85e-06 ***
## prevalentStroke  0.751412   0.483562   1.554   0.1202
## prevalentHyp     0.226231   0.135098   1.675   0.0940 .
## totChol          0.002248   0.001122   2.003   0.0452 *
## sysBP            0.014219   0.002857   4.976 6.48e-07 ***
## glucose          0.007314   0.001673   4.373 1.23e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3120.5  on 3655  degrees of freedom
## Residual deviance: 2757.2  on 3647  degrees of freedom
## AIC: 2775.2
##
## Number of Fisher Scoring iterations: 5
```

2.1.4 Visualizations

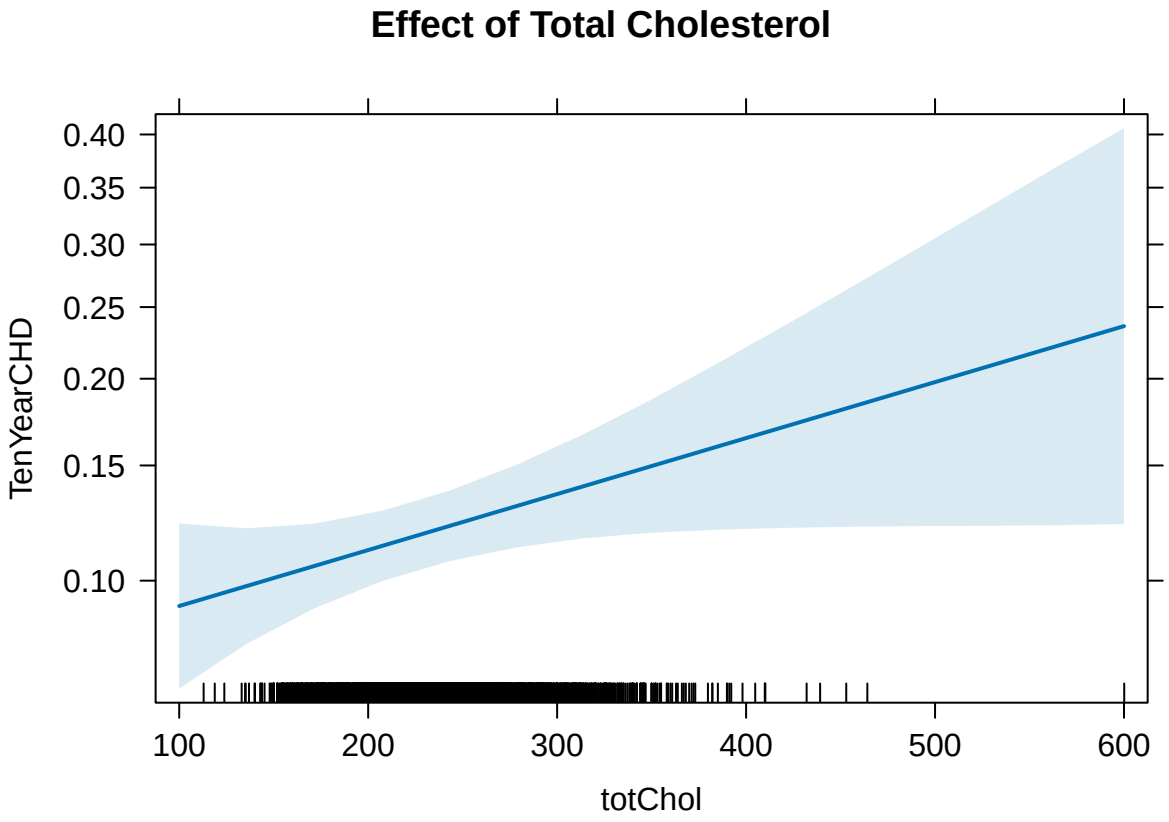
```r
# Generate and plot the effects for the final model
# Generate all effects
effects_list <- allEffects(framingham_mod_step)

# Plot each effect separately
plot(effects_list[["age"]], main = "Effect of Age on 10-Year CHD Risk")
```



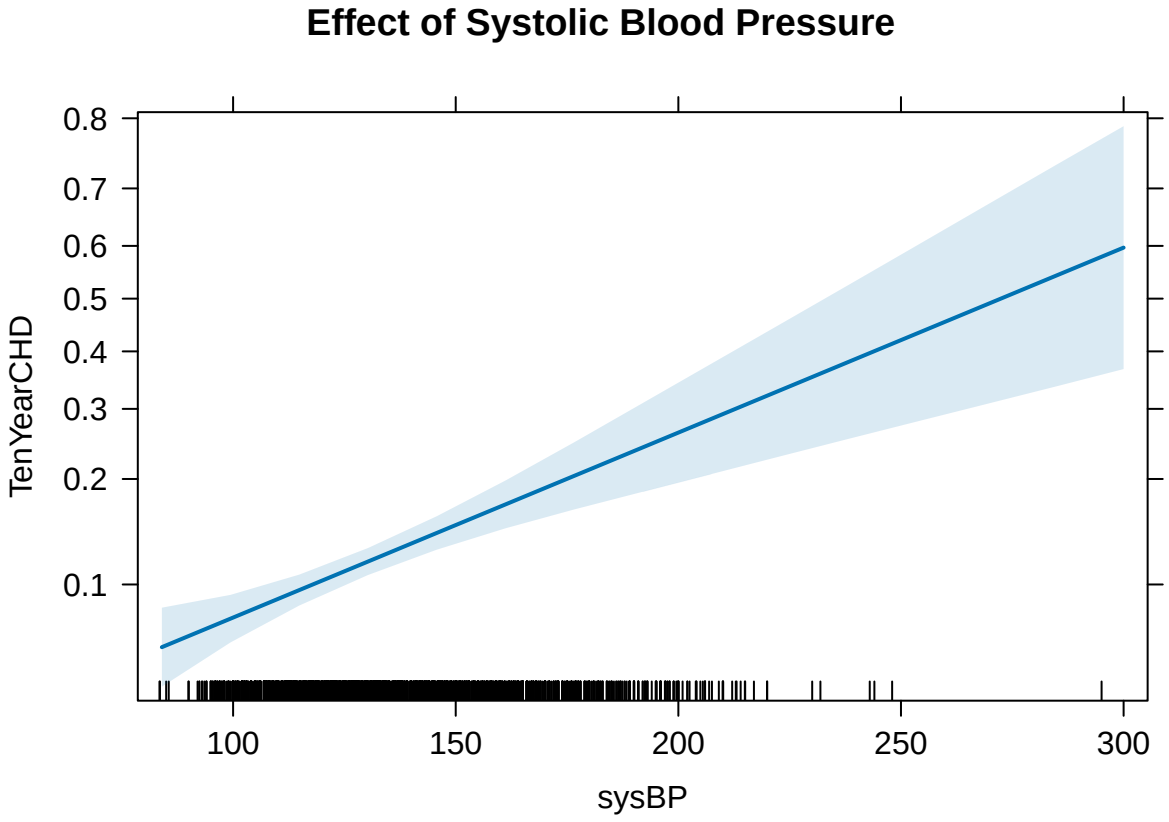

```
plot(effects_list[["cigsPerDay"]], main = "Effect of Cigarettes Per Day")
```



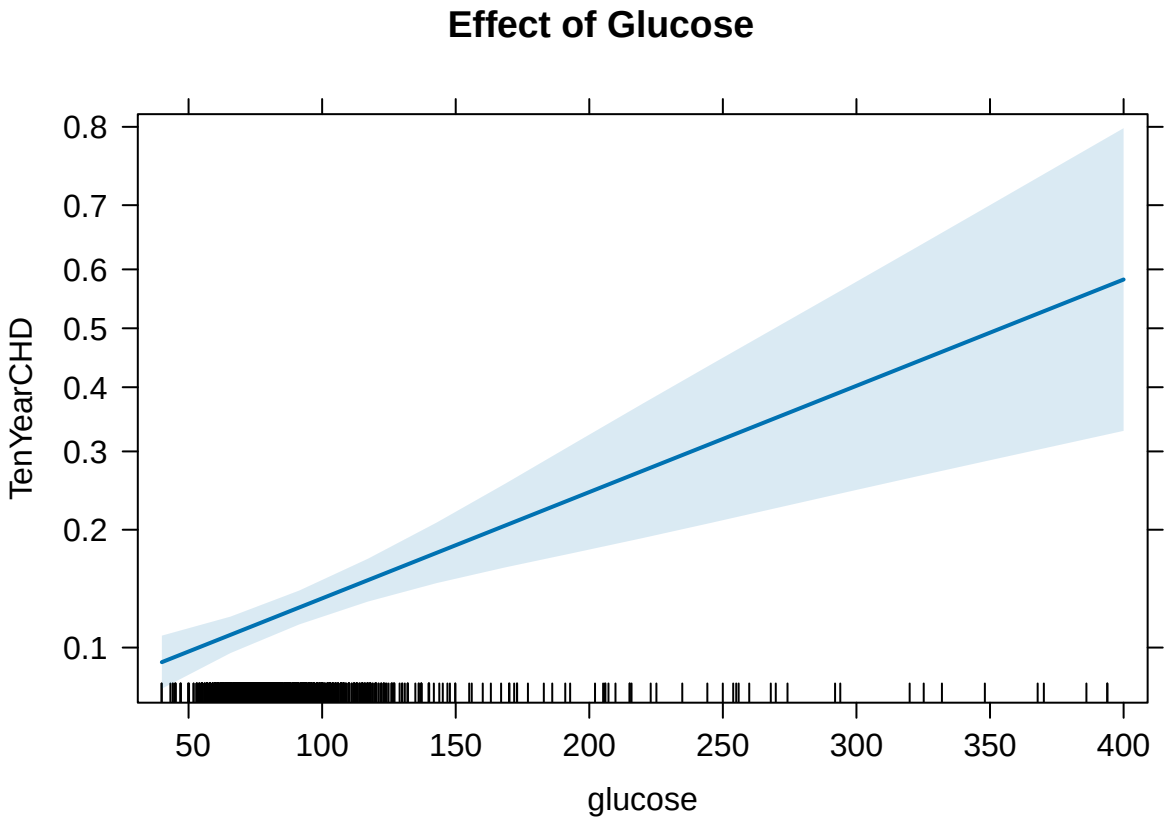

```
plot(effects_list[["totChol"]], main = "Effect of Total Cholesterol")
```




```
plot(effects_list[["sysBP"]], main = "Effect of Systolic Blood Pressure")
```




```
plot(effects_list[["glucose"]], main = "Effect of Glucose")
```



2.3 Report

Detailed Findings and Recommendations

Key Risk Factors for 10-Year CHD Risk

Based on our analysis and the final logistic regression model, the most significant risk factors contributing to the 10-year risk of coronary heart disease (CHD) include:

Age

- Older age is strongly associated with an increased CHD risk.

Cholesterol Levels (totChol)

- Higher cholesterol levels contribute significantly to CHD risk, suggesting that **lipid management** is essential.

Blood Pressure (sysBP and/or diaBP)

- Elevated blood pressure levels are associated with increased risk, highlighting the importance of **blood pressure control**.

Smoking Behavior

- The interaction between **currentSmoker** and **cigsPerDay** is significant.
  - Being a **current smoker** increases CHD risk.
  - The number of **cigarettes smoked per day** further amplifies this risk.
  - This underscores the **need for robust smoking cessation programs**.

Other Factors

- Depending on the stepwise model output, additional variables such as **diabetes status, BMI, or medication use** may also influence CHD risk.
- However, the factors listed above appear to be the most significant based on our model selection criteria.

Recommendations for Targeted Interventions

1. Age-Related Interventions

- Implement **routine cardiovascular screening programs** for older adults.
- Promote **healthy aging initiatives**, including **physical activity and diet modifications**.

2. Cholesterol Management

- Increase **community awareness** about cholesterol control.
- Encourage **regular lipid profile testing**.
- Support programs that **facilitate access to cholesterol-lowering medications** and lifestyle interventions.

3. Hypertension Control

- Organize **community blood pressure screening events**.
- Educate the public on the importance of **lifestyle changes** (e.g., **reduced salt intake, exercise**) for managing blood pressure.
- Enhance **access to antihypertensive treatment**.

4. Smoking Cessation Programs

- Develop **targeted smoking cessation campaigns**, particularly emphasizing the **compounded risk** associated with heavy smoking.
- Provide resources such as **counseling, nicotine replacement therapy, and support groups**.
- Consider interventions in **high-risk communities** identified by the model.

5. Integrated Public Health Strategies

- Utilize these insights to **inform policy decisions** and allocate resources effectively.
- Foster **collaborations between healthcare providers, local governments, and community organizations** to design interventions that address **multiple risk factors simultaneously**.

Conclusion

Our analysis of the **Framingham Heart Study** dataset using **binary logistic regression** and **stepwise model selection** has identified key **demographic and lifestyle factors**—especially **age, cholesterol levels, blood pressure, and smoking behavior**—as **critical drivers of 10-year CHD risk**.

The interaction between **smoking status** and **the number of cigarettes smoked per day** further emphasizes the **need for targeted smoking cessation efforts**. These findings provide **actionable insights** for public health officials and support the development of **targeted interventions** to reduce the burden of cardiovascular disease in our region.

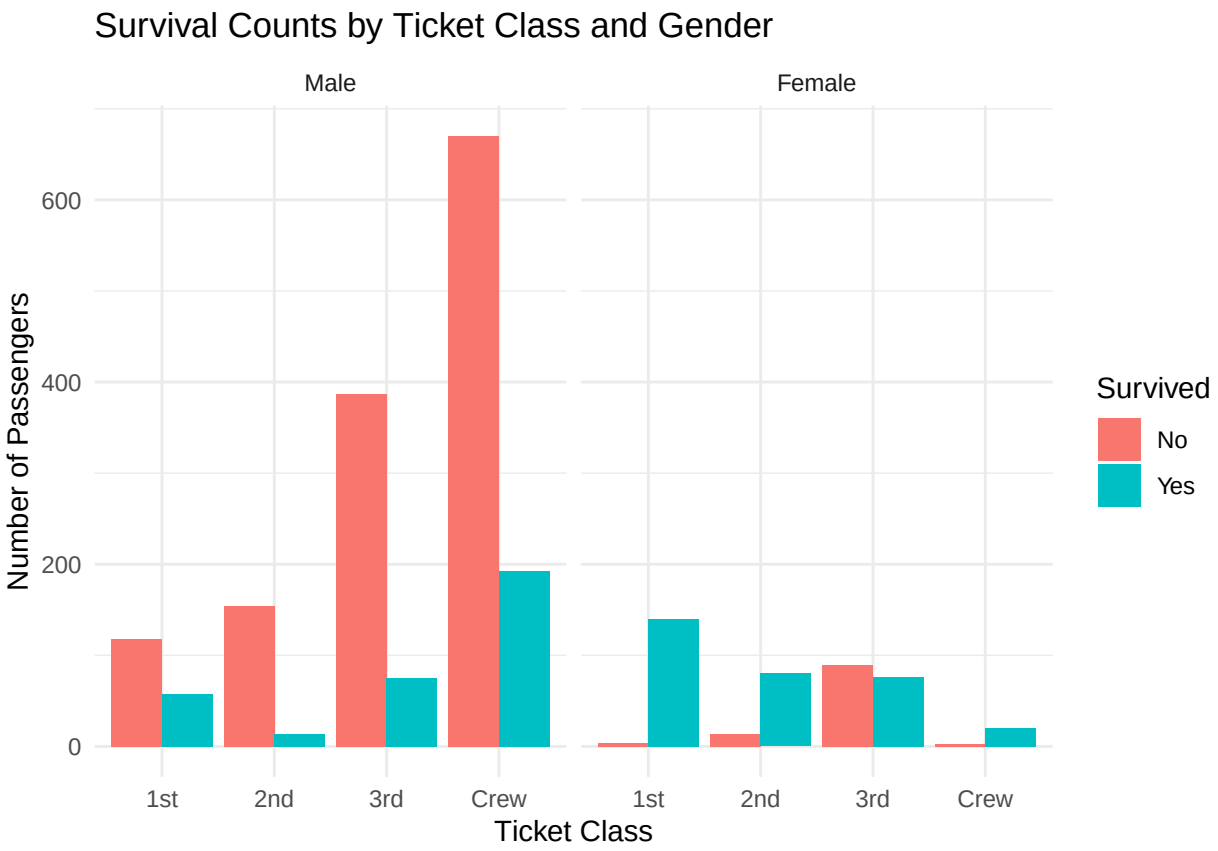
By implementing these **evidence-based strategies**, we can improve **heart health outcomes** and reduce **long-term healthcare costs** for at-risk populations.

Scenario 3: Internal Briefing Manager's Request: Your manager was very impressed with your work on the last two assignments and has asked you to present a briefing at the next team meeting. The focus is on improving team efficiency in data analysis. She asks:
"Can you provide an example of how visualization or exploratory data analysis (EDA) has helped uncover hidden trends in our datasets? Use either the Titanic or Framingham data to demonstrate this."

Your Task: 1. Select one dataset (Titanic or Framingham) and use your exploratory data analysis from the previous task. 2. Create a visualization that highlights a key trend or insight. 3. Write a short summary explaining how EDA was used to uncover this insight.

3.1 EDA and Visualization

```
# EDA and Visualization for Titanic Dataset
library(ggplot2)

# creating a bar plot to display survival counts by ticket class, with facets for gender
ggplot(titanic_df, aes(x = Class, y = Freq, fill = Survived)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ Sex) +
  labs(title = "Survival Counts by Ticket Class and Gender",
       x = "Ticket Class",
       y = "Number of Passengers") +
  theme_minimal()
```



3.2 Summary of Findings

In this visualization, the data are grouped by ticket class and separated by gender:

Gender Impact

The plot reveals a striking difference in survival counts between male and female passengers.
- Female passengers show considerably higher survival counts compared to their male counterparts across all classes.

Ticket Class Impact

Within each gender, survival outcomes also vary by ticket class.
- For **females**, survival counts are especially high in **first and second classes**.
- **Male passengers** in **third class** have the lowest survival counts.

Key Takeaway

This example of Exploratory Data Analysis (EDA) allowed our team to quickly identify that the **intersection of gender and ticket class plays a crucial role in survival outcomes**. Recognizing this pattern early on helped us to focus our subsequent modeling efforts on these key predictors, thereby improving our overall data analysis efficiency.

By leveraging such **visualizations**, we can **effectively communicate complex trends** to the team and stakeholders, ensuring that our analysis is both **data-driven and actionable**.