STAT410 Assignment 4: Generalised Linear Models

Ben Murarotto

May 2025

Contents

1 Hypertension dataset.

Does smoking or drinking change the probability of having hypertension and do these probabilities change depending on gender and/or age brackets?

Gender: gender of participant at time of study. 2 levels; 1 (Male) and 2 (Female) *Age:* Age bracket. 4 levels; 1 (20-34), 2 (35-49), 3 (50-64), 4 (>65). *Drink:* If participant had at least 12 drinks/year. 2 levels; 0 (No), 1 (Yes). ONLY in Hyper_Drink.csv *Hypertension:* Counts of participants that had either been diagnosed with hypertension or if the participant had systolic blood pressure levels $\geq$ 130 mm Hg and/or diastolic blood pressure levels $\geq$ 80 mm Hg. *Total:* Total number of participants for that Gender-Age-Cotinine/Drink combination.

To begin the analysis let's determine the proportions of participants with hypertension for each combination of gender, age, and drinking status. This provides an initial understanding of how hypertension prevalence differs across demographic groups. We will then attempt to make two visualisations to explore potential interactions.

```
drink.df <- read.csv("Hyper_Drink.csv", header = T)

drink.df$Gender <- factor(drink.df$Gender, levels = c(1, 2), labels = c("Male", "Female"))
drink.df$Age <- factor(drink.df$Age, levels = c(1, 2, 3, 4), labels = c("20-34", "35-49", "50-64", "65+"))
drink.df$Drink <- factor(drink.df$Drink, levels = c(0, 1), labels = c("Non-Drinker", "Drinker"))
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
drink.df %>%
  group_by(Gender) %>%
  summarise(
    Total = sum(Total),
    Hypertension = sum(Hypertension),
    Proportion = Hypertension / Total
  )
```

```
## # A tibble: 2 x 4
##   Gender Total Hypertension Proportion
##   <fct>  <int>        <int>      <dbl>
## 1 Male    3277         2093      0.639
## 2 Female  3385         1962      0.580
```

```
drink.df %>%
  group_by(Age) %>%
  summarise(
    Total = sum(Total),
    Hypertension = sum(Hypertension),
    Proportion = Hypertension / Total
  )
```

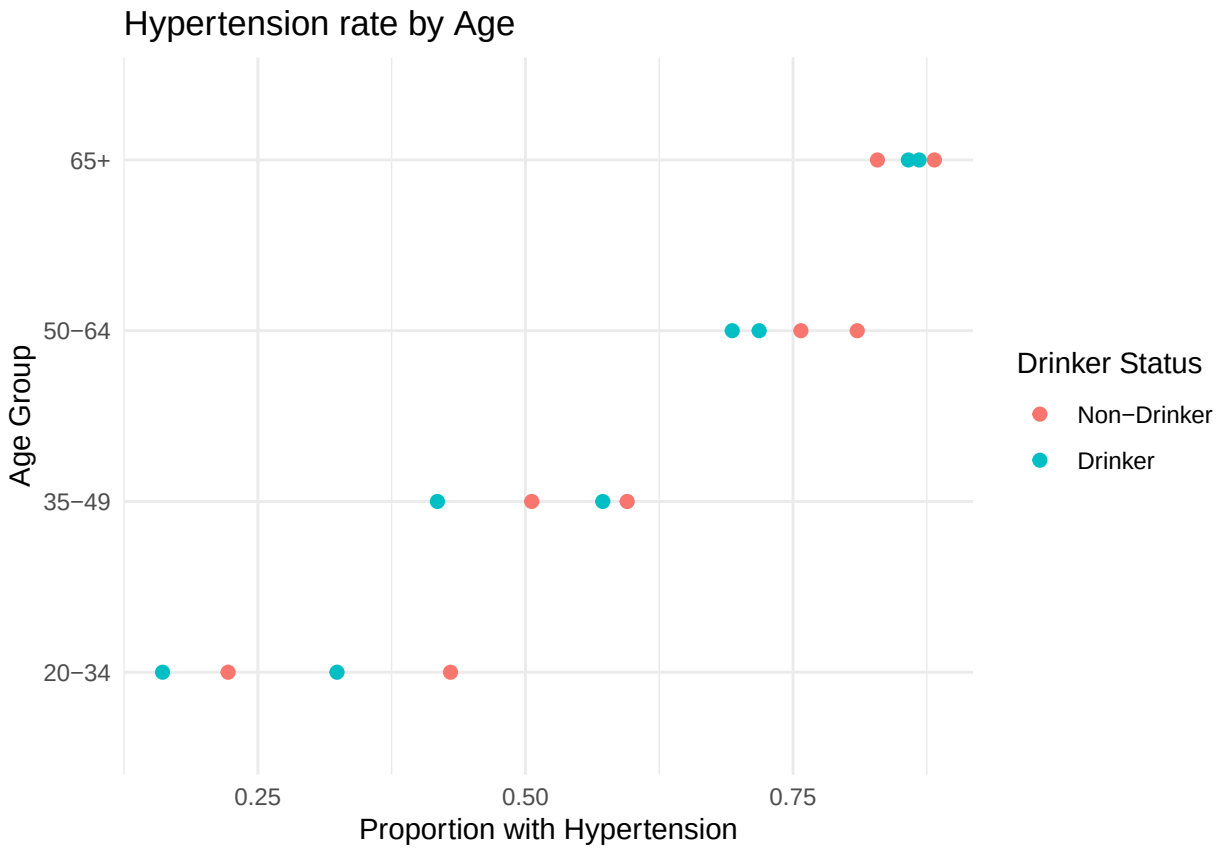
```
## # A tibble: 4 x 4
##   Age   Total Hypertension Proportion
##   <fct> <int>        <int>      <dbl>
## 1 20-34  1494          382      0.256
## 2 35-49  1590          814      0.512
## 3 50-64  1806         1328      0.735
## 4 65+    1772         1531      0.864
```

```
drink.df %>%
  group_by(Drink) %>%
  summarise(
    Total = sum(Total),
    Hypertension = sum(Hypertension),
    Proportion = Hypertension / Total
  )
```

```
## # A tibble: 2 x 4
##   Drink       Total Hypertension Proportion
##   <fct>       <int>        <int>      <dbl>
## 1 Non-Drinker  2002         1358      0.678
## 2 Drinker      4660         2697      0.579
```

```
library(ggplot2)
```

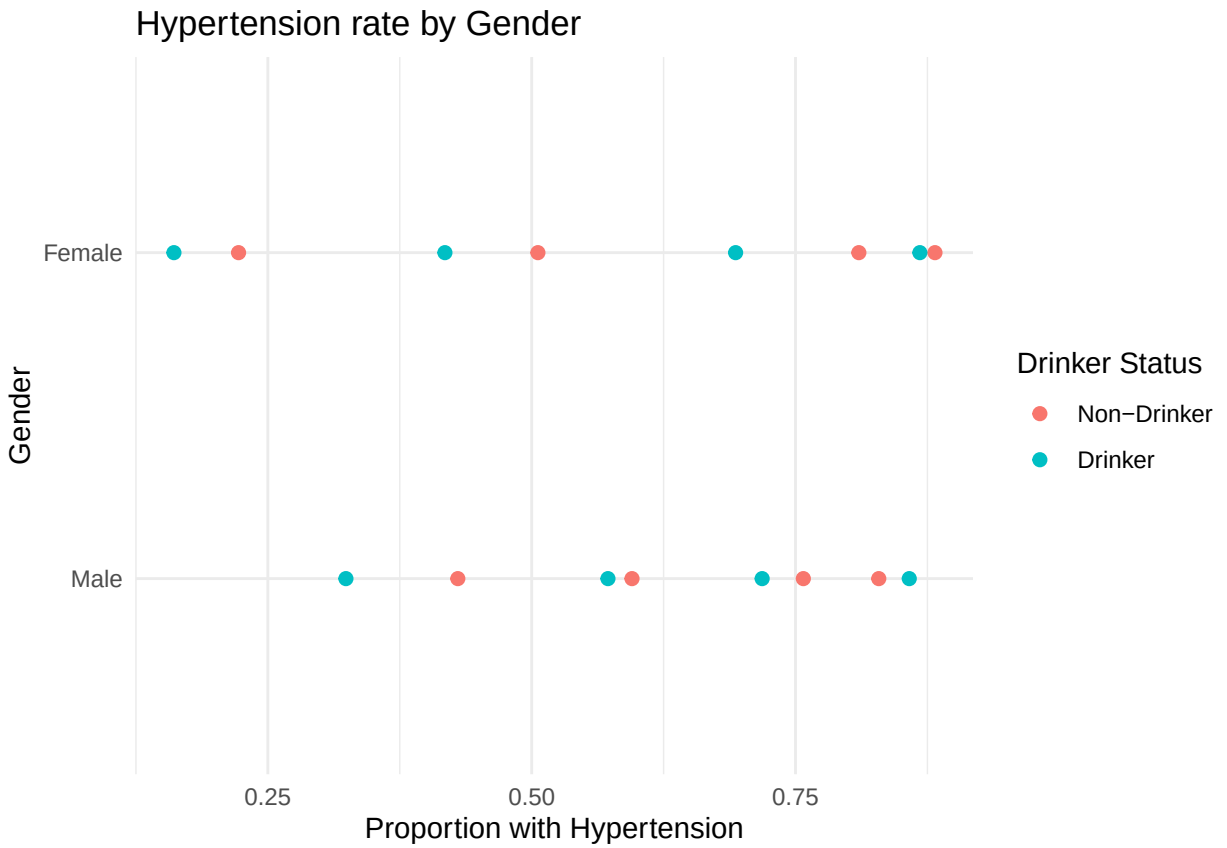
```
ggplot(drink.df, aes(
  x = Hypertension / Total,
  y = Age,
  color = Drink,
  group = Drink
)) +
  geom_point(size = 2) +
  labs(
    title = "Hypertension rate by Age",
    y = "Age Group",
    x = "Proportion with Hypertension",
    color = "Drinker Status"
  ) +
  theme_minimal()
```



The proportion of hypertension increases with age. Non drinkers show slightly higher rates of hypertension in older age groups.

```
ggplot(drink.df, aes(
  x = Hypertension / Total,
  y = Gender,
  color = Drink,
  group = Drink
)) +
  geom_point(size = 2) +
  labs(
    title = "Hypertension rate by Gender",
    y = "Gender",
    x = "Proportion with Hypertension",
```

```
    color = "Drinker Status"
  ) +
  theme_minimal()
```



Looking at both the table and the plots both genders exhibit increased hypertension with age, males generally have higher proportions of hypertension compared to females. The effect of drinking varies by gender, with a slightly stronger association in males.

1.1 Choosing a GLM.

The response variable in this research question is binomially distributed, either the patient has hypertension or does not. Therefore, a binomial family GLM plus a logit link function is appropriate.

```
model <- glm(cbind(Hypertension, Total - Hypertension) ~ Gender + Age + Drink +
               Gender:Drink + Age:Drink,
             data = drink.df,
             family = binomial(link = "logit"))

summary(model)
```

```
##
## Call:
## glm(formula = cbind(Hypertension, Total - Hypertension) ~ Gender +
##     Age + Drink + Gender:Drink + Age:Drink, family = binomial(link = "logit"),
##     data = drink.df)
##
## Coefficients:
```

```
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -0.85472    0.14527  -5.884 4.01e-09 ***
## GenderFemale               -0.12661    0.12038  -1.052   0.2929
## Age35-49                    1.07924    0.15643   6.899 5.22e-12 ***
## Age50-64                    2.29622    0.15659  14.663  < 2e-16 ***
## Age65+                      2.83025    0.16219  17.450  < 2e-16 ***
## DrinkDrinker               -0.06070    0.16324  -0.372   0.7100
## GenderFemale:DrinkDrinker -0.31567    0.13808  -2.286   0.0222 *
## Age35-49:DrinkDrinker       0.04869    0.18050   0.270   0.7874
## Age50-64:DrinkDrinker      -0.30515    0.18236  -1.673   0.0943 .
## Age65+:DrinkDrinker         0.09474    0.19729   0.480   0.6311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1631.158  on 15  degrees of freedom
## Residual deviance:   49.525  on  6  degrees of freedom
## AIC: 165.36
##
## Number of Fisher Scoring iterations: 4
```

1.2 Using stepwise regression.

Lets fit a second order interaction GLM using step() in both directions.

```
upper_model <- glm(cbind(Hypertension, Total - Hypertension) ~
                     (Gender + Age + Drink)^2,
                   data = drink.df,
                   family = binomial(link = "logit"))

final_model <- step(upper_model, direction = "both")
```

```
## Start:  AIC=123.56
## cbind(Hypertension, Total - Hypertension) ~ (Gender + Age + Drink)^2
##
##                Df Deviance    AIC
## <none>               1.725 123.56
## - Age:Drink     3    8.697 124.53
## - Gender:Drink  1    5.299 125.13
## - Gender:Age    3   49.525 165.35
```

```
summary(final_model)
```

```
##
## Call:
## glm(formula = cbind(Hypertension, Total - Hypertension) ~ (Gender +
##     Age + Drink)^2, family = binomial(link = "logit"), data = drink.df)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -0.43516    0.16065  -2.709  0.00675 **
## GenderFemale               -0.73714    0.16304  -4.521 6.15e-06 ***
## Age35-49                    0.82146    0.19278   4.261 2.03e-05 ***
## Age50-64                    1.66153    0.19418   8.557  < 2e-16 ***
```

```
## Age65+                     2.05106    0.20902   9.813  < 2e-16 ***
## DrinkDrinker              -0.27257    0.16439  -1.658  0.09730 .
## GenderFemale:Age35-49      0.37372    0.16472   2.269  0.02328 *
## GenderFemale:Age50-64      0.91802    0.16881   5.438 5.39e-08 ***
## GenderFemale:Age65+        1.11515    0.19484   5.724 1.04e-08 ***
## GenderFemale:DrinkDrinker -0.25902    0.13660  -1.896  0.05793 .
## Age35-49:DrinkDrinker      0.17711    0.18873   0.938  0.34801
## Age50-64:DrinkDrinker     -0.03529    0.19058  -0.185  0.85310
## Age65+:DrinkDrinker        0.44153    0.20819   2.121  0.03394 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1631.1581  on 15  degrees of freedom
## Residual deviance:    1.7248  on  3  degrees of freedom
## AIC: 123.56
##
## Number of Fisher Scoring iterations: 3
```

Our stepwise model kept all main effects and several interactions. When comparing coefficients of levels of our qualitative predictors we notice several things. Being female is associated with significantly lower odds of hypertension. Age is a strong predictor, with risk of hypertension increasing with age. Interaction effects confirm our suspicions that the influence of drinking and age on hypertension differs between genders and across age groups.

1.3 Understanding residual deviance.

```
anova(final_model, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(Hypertension, Total - Hypertension)
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                            15    1631.16
## Gender        1    24.42        14    1606.73 7.733e-07 ***
## Age           3  1521.77        11      84.97 < 2.2e-16 ***
## Drink         1    23.75        10      61.21 1.095e-06 ***
## Gender:Age    3    48.79         7      12.42 1.443e-10 ***
## Gender:Drink  1     3.72         6       8.70   0.05369 .
## Age:Drink     3     6.97         3       1.72   0.07280 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pchisq(df=3, q=1.7248, lower.tail = F)
```

```
## [1] 0.6314343
```

The hypothesis testing for the chi-square test is as follows:

H0: The model fits well (no lack of fit).

HA: The model does not fit well.

The p-value 0.631 for the chi square test in this instance we fail to reject the null hypothesis and accept that the model is a good fit.

1.4 Final summary.

Drinking has a notable influence on the likelihood of having hypertension. In the analysis of deviance, the main effect of drinking was statistically significant (Deviance = 23.75, p = 1.095 × 10e-6), showing that drinking status contributes meaningfully to the model. Additionally, its impact isn't necessarily consistent across the board, the effect of drinking changes depending on both age and gender. This is reflected in the retention of the Age:Drink interaction (Deviance = 6.97, p = 0.073) and the Gender:Drink interaction (Deviance = 3.72, p = 0.054) in the final stepwise selected model. While these interactions are only marginally significant, their inclusion suggests that drinking affects hypertension risk differently for different age and gender groups.

2 Diabates dataset.

1.2 million adults in USA are diagnosed with diabetes each year *(American Diabetes Association 2025)*. Complications can arise in a number of organ systems due to diabetes, severely impacting quality of life. Researchers want to understand what some important risk factors of diabetes are. The research questions is: How does gender, age, income, BMI and other disorders influence the probability of having diabetes? The data is stored in Diabetes.csv and consists of variables: - *Seqn*: Unique sequence number from participant. - *Gender*: gender of participant at time of study. 2 levels; 1 (Male) and 2 (Female) - *Age*: Age bracket. 4 levels; 1 (20-34), 2 (35-49), 3 (50-64), 4 (>65). - *PIR*: Ratio of family income to poverty. Calculated by dividing family (or individual) income by the poverty threshold specific to each survey year. Values above 1 indicate the individual is above the poverty threshold and values below 1 indicate they are below the poverty threshold. - *BMI*: Body mass index. 3 levels; 1 (<25, normal weight), 2 (25-30, overweight), 3 (<30, obese). - *Diabetes*: Whether the participant had been diagnosed with diabetes or if the participant had high glucose. - *Nocturia*: Whether the participant was deemed to experience nocturia. - *Hypertension*: Whether the participant had been diagnosed with hypertension or elevated BP.

```
diabetes.df <- read.csv("Diabetes.csv", header = T)
diabetes.df$Gender <- factor(diabetes.df$Gender, levels = c(1, 2), labels = c("Male", "Female"))
diabetes.df$Age <- factor(diabetes.df$Age, levels = c(1, 2, 3, 4), labels = c("20-34", "35-49", "50-64", "6
diabetes.df$BMI <- factor(diabetes.df$BMI, levels = c(1, 2, 3), labels = c("Normal", "Overweight", "Obese")
diabetes.df$Nocturia <- factor(diabetes.df$Nocturia, levels = c(0, 1), labels = c("No", "Yes"))
diabetes.df$Hypertension <- factor(diabetes.df$Hypertension, levels = c(0, 1), labels = c("No", "Yes"))


set.seed(7)
sample <- diabetes.df%>%sample_n(200)
df <- subset(sample, select = -seqn)
db.glm <- glm(data = df, Diabetes~., family="binomial"(link="logit"))
summary(db.glm)
```

```
##
## Call:
## glm(formula = Diabetes ~ ., family = binomial(link = "logit"),
##     data = df)
##
## Coefficients:
```

```
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.4235     0.9045  -3.785 0.000154 ***
## GenderFemale     -0.6238     0.3924  -1.590 0.111935
## Age35-49          0.8033     0.7636   1.052 0.292814
## Age50-64          1.4013     0.7391   1.896 0.057946 .
## Age65+            2.3149     0.7477   3.096 0.001962 **
## PIR              -0.1493     0.1252  -1.192 0.233115
## BMIOverweight     0.3585     0.6300   0.569 0.569305
## BMIObese          1.7840     0.6321   2.822 0.004765 **
## NocturiaYes       0.9279     0.3870   2.398 0.016506 *
## HypertensionYes   0.6251     0.4967   1.259 0.208197
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 246.02  on 199  degrees of freedom
## Residual deviance: 177.94  on 190  degrees of freedom
## AIC: 197.94
##
## Number of Fisher Scoring iterations: 5
```

$$\log\left(\frac{p}{1-p}\right) = -3.42 - 0.62 \cdot \text{Gender}_{\text{Female}} + 0.80 \cdot \text{Age}_{35-49} + 1.40 \cdot \text{Age}_{50-64} + 2.31 \cdot \text{Age}_{65+} - 0.15 \cdot \text{PIR} + 0.36 \cdot \text{BMI}_{\text{Overweight}} + 1.78 \cdot \text{BMI}_{\text{Obese}} + 0.93 \cdot \text{Nocturia}_{\text{Yes}} + 0.63 \cdot \text{Hypertension}_{\text{Yes}}$$

```
anova(db.glm, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Diabetes
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         199     246.02
## Gender        1    0.581       198     245.44  0.445791
## Age           3   33.241       195     212.19 2.864e-07 ***
## PIR           1    4.605       194     207.59  0.031872 *
## BMI           2   20.797       192     186.79 3.049e-05 ***
## Nocturia      1    7.239       191     179.55  0.007136 **
## Hypertension  1    1.611       190     177.94  0.204291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pchisq(df=190, q=177.94, lower.tail = F)
```

```
## [1] 0.7250365
```

Like in our previous dataset, the goodness of fit test p-value is large therefore we do not reject our null hypothesis and we can say that this model is a good fit for the data.

2.1 Determining odds ratio.

Let's calculate the CI range of odds ratios for each of the predictors of the main effects model. This will give us a better understanding of our predictor weights.

```
odds <- exp(coef(db.glm))
odds.conf <- exp(confint(db.glm))
```

```
## Waiting for profiling to be done...
```

```
odds.df <- data.frame(
  CI_Lower = round(odds.conf[, 1], 3),
  Odds_Ratio = round(odds, 3),
  CI_Upper = round(odds.conf[, 2], 3)
)

odds.df
```

```
##                 CI_Lower Odds_Ratio CI_Upper
## (Intercept)        0.005      0.033    0.167
## GenderFemale       0.244      0.536    1.145
## Age35-49           0.536      2.233   11.651
## Age50-64           1.047      4.061   20.542
## Age65+             2.604     10.123   52.352
## PIR                0.671      0.861    1.099
## BMIOverweight      0.432      1.431    5.311
## BMIObese           1.845      5.954   22.686
## NocturiaYes        1.194      2.529    5.487
## HypertensionYes    0.713      1.868    5.090
```

When interpreting CI intervals in logistic regression, we remember that it indicates the range of values within which the true odds ratio is likely to fall if the CI excludes 1, the effect is considered statistically significant.

Among all significant predictors, older age 65+, obesity, and nocturia were significantly associated with increased odds of diabetes. Age 65+ had the strongest effect, with participants in this age group experiencing 10 times higher odds compared 20–34 age group.

2.2 Making predictions.

Let's now use our main effects model to make a prediction with some new data.

```
newdata.pred <- data.frame(
  Gender = factor("Male", levels = levels(df$Gender)),
  Age = factor("35-49", levels = levels(df$Age)),
  PIR = 1.7,
  BMI = factor("Normal", levels = levels(df$BMI)),
  Nocturia = factor("Yes", levels = levels(df$Nocturia)),
  Hypertension = factor("Yes", levels = levels(df$Hypertension))
)
```

```
pred <- predict(db.glm, newdata = newdata.pred, type = "link", se.fit = TRUE)

fit <- pred$fit
lwr <- fit - 1.96 * pred$se.fit
upr <- fit + 1.96 * pred$se.fit
```

```r
prob <- plogis(fit)
lower_prob <- plogis(lwr)
upper_prob <- plogis(upr)

range <- c(Probability = prob, Lower_CI = lower_prob, Upper_CI = upper_prob)
range
```

```
## Probability.1    Lower_CI.1    Upper_CI.1
##    0.21066745    0.05996247    0.52756879
```

With 95% confindence we can say that the probability of a 35-49 year old male with hypertension, nocturia and a PIR of 1.7 and normal BMI lies between 6% and 52%.

2.3 Visualisiing PIR and Gender.

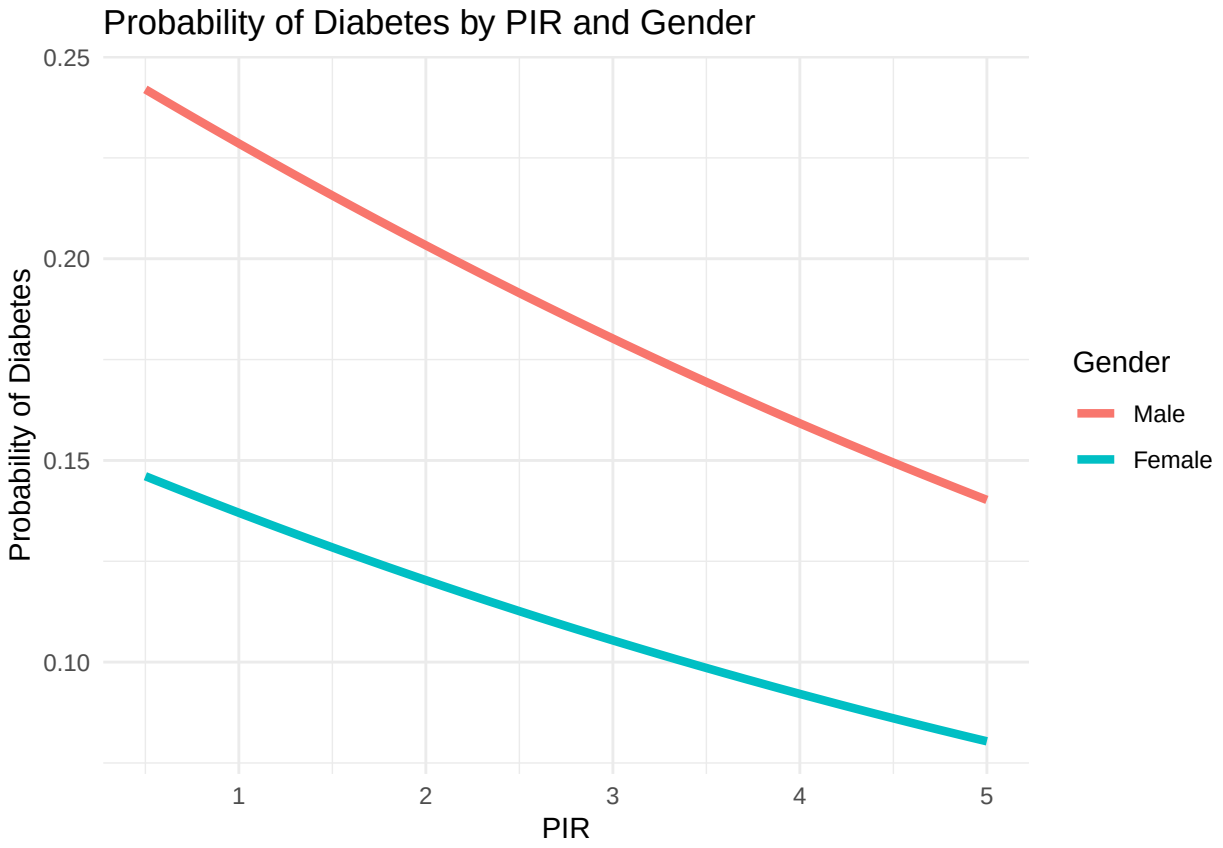
Let's visualise how gender and poverty affect diabetes risk. We need to create a set of predictions of diabetes using the fitted GLM across a range of PIR and then compare the genders.

```r
newdata <- expand.grid(
  Gender = factor(c("Male", "Female"), levels = levels(df$Gender)),
  Age = factor("35-49", levels = levels(df$Age)),
  PIR = seq(0.5, 5, 0.05),
  BMI = factor("Normal", levels = levels(df$BMI)),
  Nocturia = factor("Yes", levels = levels(df$Nocturia)),
  Hypertension = factor("Yes", levels = levels(df$Hypertension))
)

newdata$fit_prob <- predict(db.glm, newdata = newdata, type = "response")


ggplot(newdata, aes(x = PIR, y = fit_prob, color = Gender)) +
  geom_line(size = 1.5) +
  labs(
    title = "Probability of Diabetes by PIR and Gender",
    x = "PIR",
    y = "Probability of Diabetes",
    color = "Gender"
  ) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



Our plot shows a clear negative relationship between PIR and probability of diabetes. People with higher PIR earn more relative to the poverty line tend to have a lower risk of diabetes. This pattern holds for both men and women, although men consistently have higher predicted probabilities of diabetes than women at every PIR level. Overall, the results suggest that socioeconomic status, as measured by PIR, plays an important role in diabetes risk.