# 7 Generalized Linear Models

Immanuel Klein

```
library("tidyverse")
library("ggplot2")
```

This task is about analyzing the student-mat.csv dataset to identify variables that explain mathematics grades (`G1`, `G2`, `G3`). First, we assess the distribution of each grade using graphical tools and statistical tests to see whether they follow a normal or a Poisson distribution. We also check for over-dispersion and other anomalies. Then, we fit a GLM to explain `G1` using all explanatory variables (Model 1), assessing the significance of covariates and the model's goodness-of-fit through residual analysis. Next, we create a reduced model (Model 2) by selecting key covariates, and compare its performance to Model 1 using an analysis of deviance test. Lastly, we modify Model 2 to replace one variable (`goout`) with another (`Walc`), which creates Model 3, and compare the models to determine which has the best explanation for `G1`.

```
students.math <- read.csv("student-mat.csv", sep = ",")
head(students.math)
```

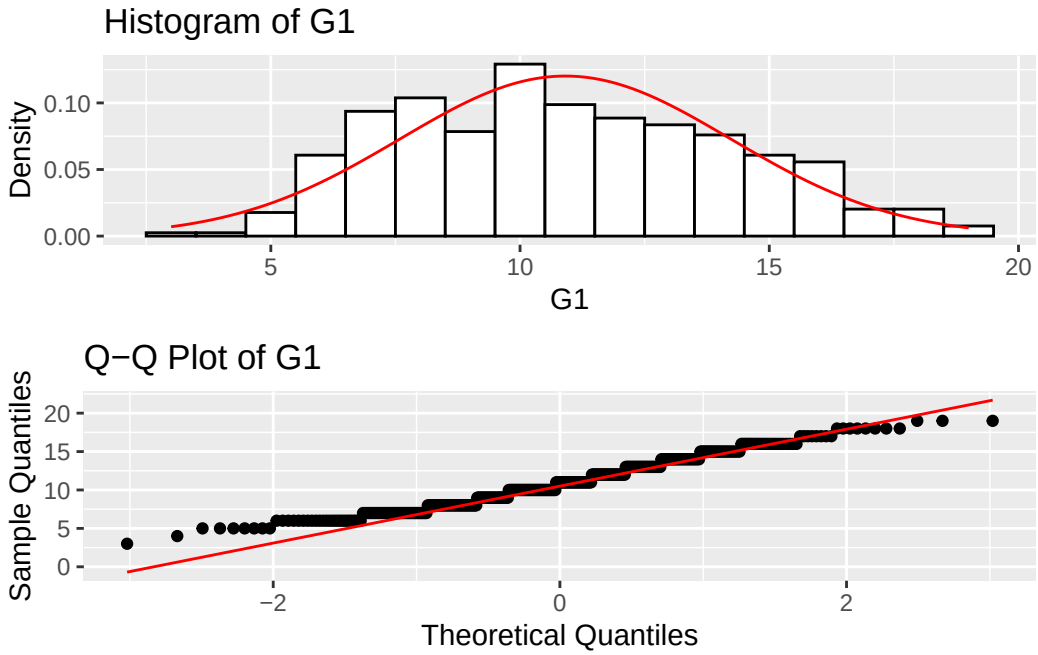
```
  school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
2     GP   F  17       U     GT3       T    1    1  at_home    other     course
3     GP   F  15       U     LE3       T    1    1  at_home    other      other
4     GP   F  15       U     GT3       T    4    2   health services       home
5     GP   F  16       U     GT3       T    3    3    other    other       home
6     GP   M  16       U     LE3       T    4    3 services    other reputation
  guardian traveltime studytime failures schoolsup famsup paid activities
1   mother          2         2        0       yes     no   no         no
2   father          1         2        0        no    yes   no         no
3   mother          1         2        3       yes     no  yes         no
4   mother          1         3        0        no    yes  yes        yes
5   father          1         2        0        no    yes  yes         no
6   mother          1         2        0        no    yes  yes        yes
```

```
  nursery higher internet romantic famrel freetime goout Dalc Walc health
1     yes    yes       no       no      4        3     4    1    1      3
2      no    yes      yes       no      5        3     3    1    1      3
3     yes    yes      yes       no      4        3     2    2    3      3
4     yes    yes      yes      yes      3        2     2    1    1      5
5     yes    yes       no       no      4        3     2    1    2      5
6     yes    yes      yes       no      5        4     2    1    2      5
  absences G1 G2 G3
1        6  5  6  6
2        4  5  5  6
3       10  7  8 10
4        2 15 14 15
5        4  6 10 10
6       10 15 15 15
```

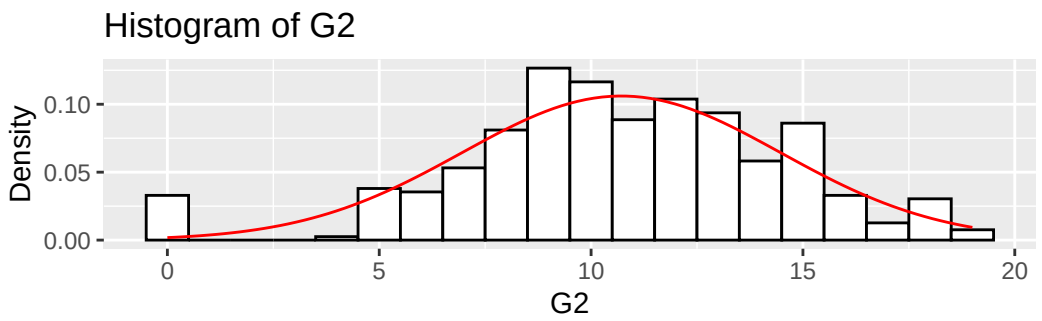
### Exercise (a)

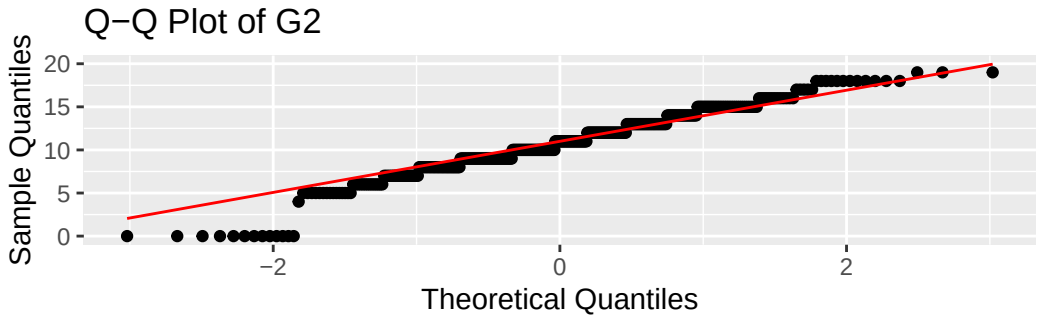
```r
# Load necessary libraries
library(ggplot2)
library(readr)

# Function to create histogram and Q-Q plot
plot.distribution <- function(data, data.name) {
  hist.plot <- ggplot(data.frame(data), aes(x = data)) +
    geom_histogram(aes(y = after_stat(density)),
                   binwidth = 1,
                   fill = "white",
                   color = "black") +
    stat_function(fun = dnorm,
                  args = c(mean = mean(data), sd = sd(data)),
                  color = "red") +
    labs(title = paste("Histogram of", data.name),
         x = data.name,
         y = "Density")

  qq.plot <- ggplot(data.frame(data), aes(sample = data)) +
    stat_qq() +
    stat_qq_line(color = "red") +
    labs(title = paste("Q-Q Plot of", data.name),
         x = "Theoretical Quantiles",
         y = "Sample Quantiles")
```

```r
  gridExtra::grid.arrange(hist.plot, qq.plot)
}

# Poisson distribution check
# Compare mean and variance for Poisson assumption
mean.var.disp <- function(variable, variable.name) {
  cat("Mean, variance, and dispersion for", variable.name, "\n")
  cat("Mean:", mean(variable), "\n")
  cat("Variance:", var(variable), "\n")

  # Check for overdispersion
  fit <- glm(variable ~ 1, family = poisson)
  dispersion <- sum(residuals(fit, type = "pearson")^2) /
    fit$df.residual
  cat(paste("Dispersion:", dispersion, "\n\n"))
}

# Plot distributions for G1, G2, and G3
plot.distribution(students.math$G1, "G1")
```

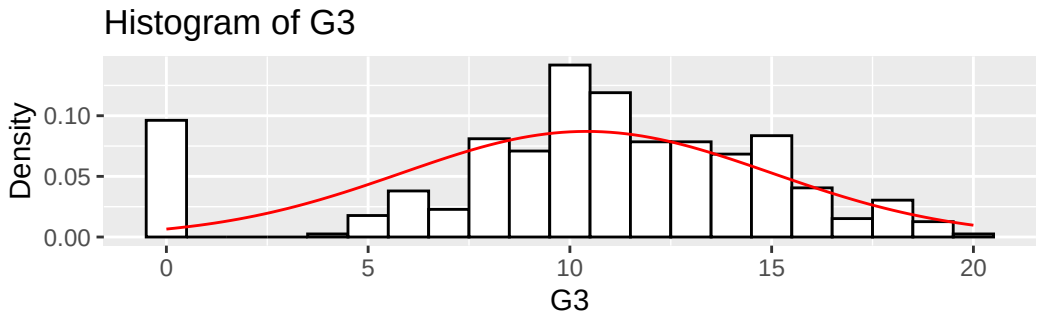


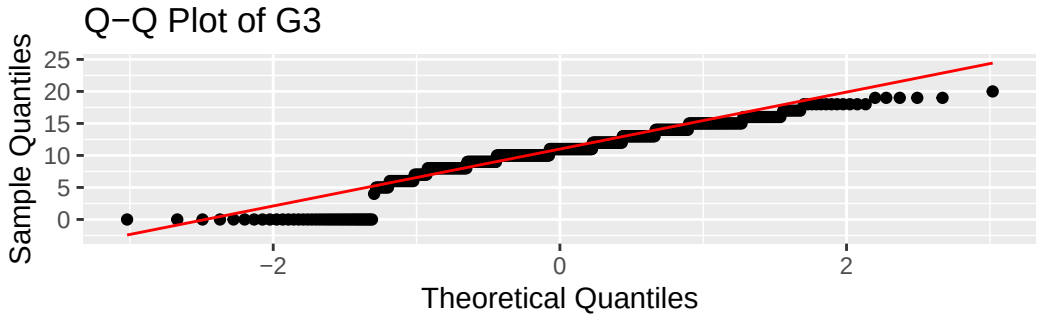
```
plot.distribution(students.math$G2, "G2")
```




```
plot.distribution(students.math$G3, "G3")
```



```r
# Statistical tests for normality
shapiro.test(students.math$G1)
```

```
    Shapiro-Wilk normality test

data:  students.math$G1
W = 0.97491, p-value = 2.454e-06
```

```r
shapiro.test(students.math$G2)
```

```
    Shapiro-Wilk normality test

data:  students.math$G2
W = 0.96914, p-value = 2.084e-07
```

```r
shapiro.test(students.math$G3)
```

```
    Shapiro-Wilk normality test

data:  students.math$G3
W = 0.92873, p-value = 8.836e-13
```

```r
# Poisson
mean.var.disp(students.math$G1, "G1")
```

```
Mean, variance, and dispersion for G1
Mean: 10.90886
Variance: 11.01705
Dispersion: 1.00991785579217
```

```r
mean.var.disp(students.math$G2, "G2")
```

```
Mean, variance, and dispersion for G2
Mean: 10.71392
Variance: 14.14892
Dispersion: 1.32061019257731
```

```r
mean.var.disp(students.math$G3, "G3")
```

```
Mean, variance, and dispersion for G3
Mean: 10.41519
Variance: 20.98962
Dispersion: 2.01528888459851
```

- Normality: If the histograms and Q-Q plots show deviations from a bell curve and the Shapiro-Wilk test has p-values $< 0.05$, the grades do not follow a normal distribution. This is the case here. `G1` performs best but still not good enough.
- Poisson Distribution: If the mean and variance are not close, and the dispersion value is significantly greater than 1, the data does not follow a Poisson distribution and may show signs of over-dispersion.
  - G1: Fits a Poisson distribution quite well. `G1` could be reasonably approximated by a Poisson distribution.
  - G2: Shows slight over-dispersion, indicating a mild deviation from a Poisson distribution.
  - G3: Exhibits significant over-dispersion, indicating that a Poisson distribution is not appropriate.

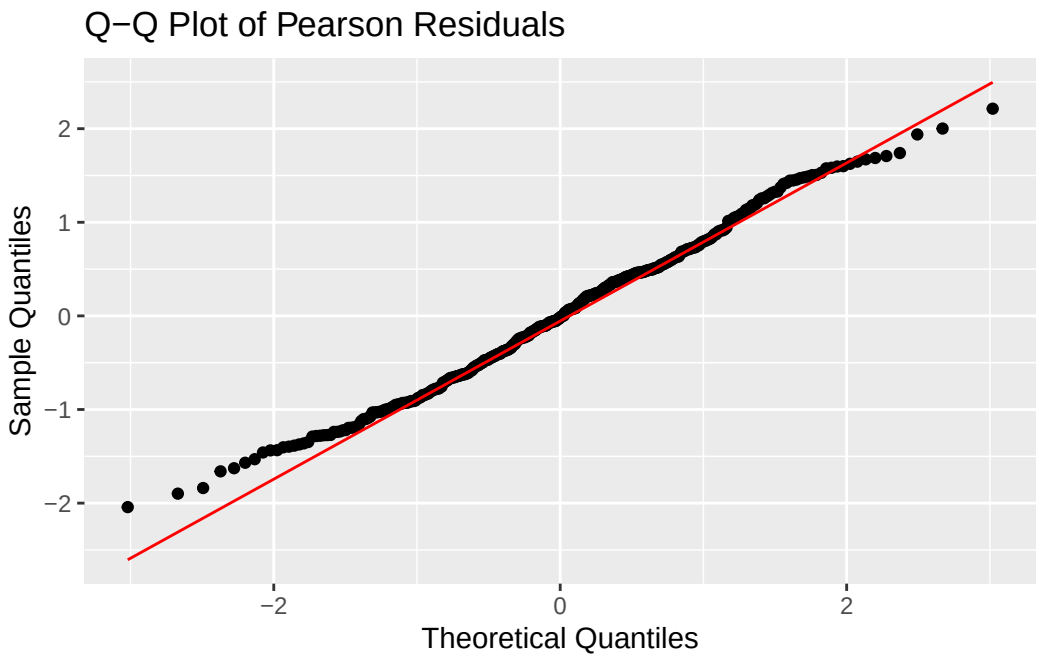
### Exercise (b)

```r
# Function to create ha Q-Q plot for residuals
plot.residuals <- function(residuals, residual.name) {
  ggplot(data.frame(residuals), aes(sample = residuals)) +
    stat_qq() +
    stat_qq_line(color = "red") +
    labs(title = paste("Q-Q Plot of", residual.name),
         x = "Theoretical Quantiles",
         y = "Sample Quantiles")
}

# Model 1: Using all explanatory variables
model1 <- glm(G1 ~ . - G2 - G3,
              data = students.math,
              family = poisson())
summary(model1)
```

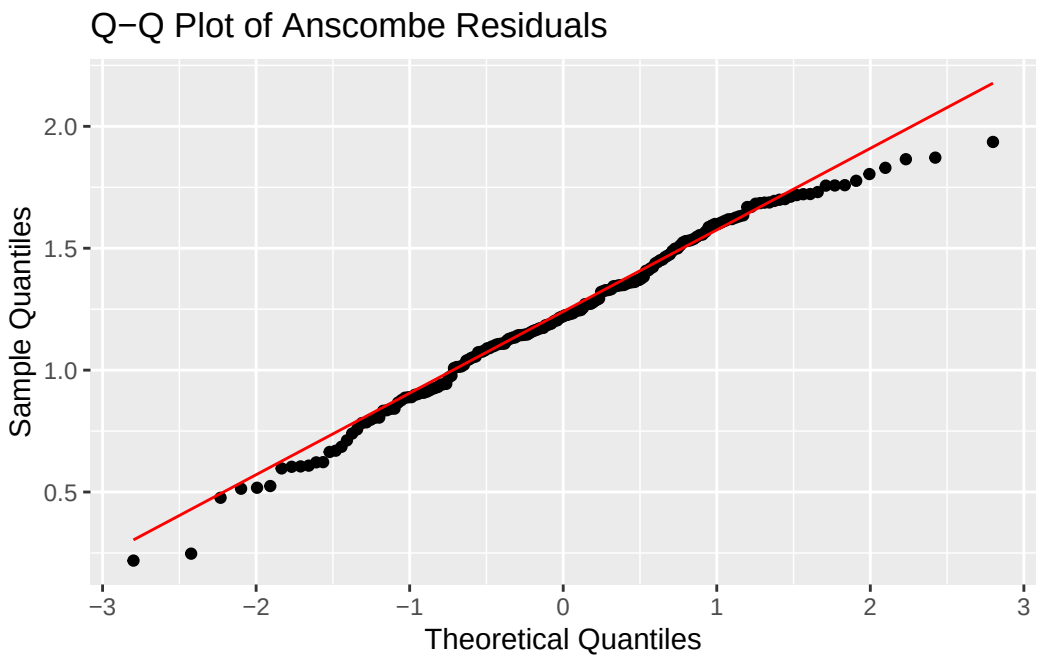
```
Call:
glm(formula = G1 ~ . - G2 - G3, family = poisson(), data = students.math)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.418107   0.331755   7.289 3.13e-13 ***
schoolMS          0.007157   0.058426   0.122  0.90250
sexM              0.078237   0.036434   2.147  0.03176 *
age              -0.006480   0.015983  -0.405  0.68515
addressU          0.012823   0.043630   0.294  0.76882
famsizeLE3        0.036639   0.035590   1.029  0.30325
PstatusT          0.016047   0.053050   0.302  0.76227
Medu              0.010664   0.023797   0.448  0.65405
Fedu              0.014886   0.020349   0.732  0.46446
Mjobhealth        0.078652   0.081444   0.966  0.33419
Mjobother        -0.077025   0.054363  -1.417  0.15653
Mjobservices      0.043696   0.059492   0.734  0.46266
Mjobteacher      -0.083899   0.076545  -1.096  0.27305
Fjobhealth       -0.042222   0.103856  -0.407  0.68434
Fjobother        -0.100107   0.073860  -1.355  0.17530
Fjobservices     -0.086835   0.076152  -1.140  0.25416
Fjobteacher       0.096068   0.091322   1.052  0.29281
reasonhome        0.015945   0.041353   0.386  0.69982
reasonother      -0.017291   0.060594  -0.285  0.77537
reasonreputation  0.038800   0.042441   0.914  0.36061
guardianmother   -0.002677   0.039862  -0.067  0.94645
guardianother     0.091135   0.074209   1.228  0.21942
traveltime       -0.003407   0.025788  -0.132  0.89488
studytime         0.053887   0.020846   2.585  0.00974 **
failures         -0.147644   0.028151  -5.245 1.56e-07 ***
schoolsupyes     -0.211886   0.052610  -4.027 5.64e-05 ***
famsupyes        -0.093311   0.035213  -2.650  0.00805 **
paidyes          -0.008676   0.035007  -0.248  0.80426
activitiesyes    -0.007168   0.032921  -0.218  0.82763
nurseryyes        0.004504   0.041064   0.110  0.91265
higheryes         0.123398   0.087908   1.404  0.16041
internetyes       0.019524   0.046777   0.417  0.67639
romanticyes      -0.018513   0.034877  -0.531  0.59555
famrel            0.001997   0.018196   0.110  0.91259
freetime          0.023120   0.017435   1.326  0.18482
goout            -0.037487   0.016743  -2.239  0.02516 *
Dalc             -0.002047   0.025074  -0.082  0.93494
```

```
Walc            -0.004639   0.018663  -0.249  0.80371
health          -0.015492   0.011868  -1.305  0.19178
absences         0.001520   0.002185   0.696  0.48662
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 402.47  on 394  degrees of freedom
Residual deviance: 263.08  on 355  degrees of freedom
AIC: 2000.2

Number of Fisher Scoring iterations: 4
```

```r
# Calculate Pearson residuals
pearson.residuals <- residuals(model1, type = "pearson")

# Calculate Anscombe residuals
anscombe.residuals <- residuals(model1, type = "response")^
  (1/3)

# Plot Pearson residuals
plot.residuals(pearson.residuals, "Pearson Residuals")
```



```r
# Plot Anscombe residuals
plot.residuals(anscombe.residuals, "Anscombe Residuals")
```



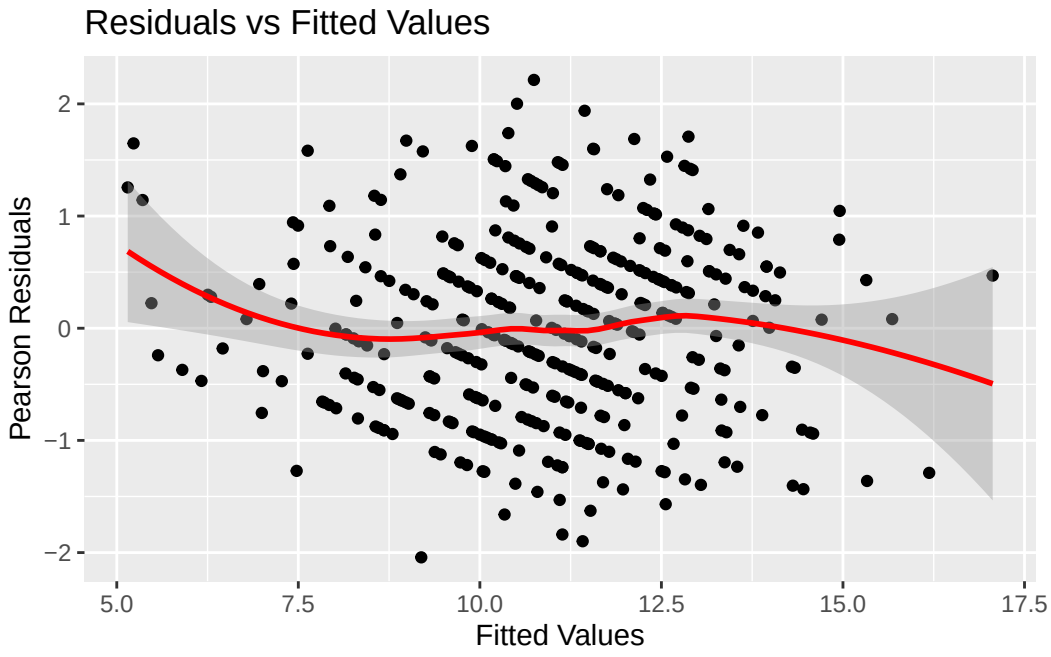

```r
# Shapiro-Wilk test for normality of residuals
shapiro.test(pearson.residuals)
```

```
    Shapiro-Wilk normality test

data:  pearson.residuals
W = 0.99256, p-value = 0.04651
```

```r
shapiro.test(anscombe.residuals)
```

```
    Shapiro-Wilk normality test

data:  anscombe.residuals
W = 0.98762, p-value = 0.08701
```

```r
# Residuals vs Fitted values plot
ggplot(data.frame(fitted = fitted(model1),
                  residuals = pearson.residuals),
       aes(x = fitted, y = residuals)) +
  geom_point() +
  geom_smooth(method = "loess", col = "red") +
  labs(title = "Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Pearson Residuals")
```




**Significance of Covariates**

- `failures` and `schoolsupyes`: Highly significant with p-values $< 0.001$.
- `studytime` and `famsupyes`: Very significant with p-values between 0.001 and 0.01.
- `goout` and `sexM`: Significant with p-values between 0.01 and 0.05.
- Other Covariates: Not significant with p-values  0.1.

**Goodness-of-Fit**

- The AIC value of this model is 2000.2.

- Without comparison to other models (e.g., models 2 and 3), we cannot make definitive statements about the goodness-of-fit. The absolute AIC value alone is not sufficient for interpretation.

**Residual Analysis**

- Q-Q Plots: For both Pearson and Anscombe residuals, there is noticeable deviation from the diagonal in the tails, indicating potential issues with normality.
- Shapiro-Wilk Test:
  - Pearson residuals: p-value = 0.04651, suggesting deviation from normality ($p < 0.05$).
  - Anscombe residuals: p-value = 0.08701, suggesting they follow a normal distribution ($p > 0.05$).
- Residuals vs Fitted Values Plot: Residuals are fairly randomly distributed with no clear pattern, though they thin out at the upper and lower ends of the range.

**Conclusion**

- The residuals are somewhat close to following a normal distribution, but there are noticeable deviations.
- The GLM appears to be moderately adequate for the data, though it is not a perfect fit. There is room for improvement in the model to better capture the data's underlying structure.

## Exercise (c)

```r
# Model 2: GLM with reduced covariates
model2 <- glm(G1 ~ sex +
                Fedu +
                studytime +
                failures +
                schoolsup +
                famsup +
                goout,
              data = students.math,
              family = poisson())
summary(model2)
```

```
Call:
glm(formula = G1 ~ sex + Fedu + studytime + failures + schoolsup +
    famsup + goout, family = poisson(), data = students.math)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.34585    0.07578  30.955  < 2e-16 ***
sexM          0.06585    0.03237   2.034  0.04193 *
Fedu          0.04281    0.01482   2.888  0.00388 **
studytime     0.05828    0.01906   3.057  0.00223 **
failures     -0.13876    0.02495  -5.561 2.69e-08 ***
schoolsupyes -0.19834    0.04978  -3.984 6.78e-05 ***
famsupyes    -0.07330    0.03240  -2.263  0.02365 *
goout        -0.03525    0.01406  -2.506  0.01220 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 402.47  on 394  degrees of freedom
Residual deviance: 302.02  on 387  degrees of freedom
AIC: 1975.1

Number of Fisher Scoring iterations: 4
```

```r
# Analysis of deviance test between Model 1 and Model 2
deviance.analysis <- anova(model1, model2, test = "Chi")
print(deviance.analysis)
```

```
Analysis of Deviance Table

Model 1: G1 ~ (school + sex + age + address + famsize + Pstatus + Medu +
    Fedu + Mjob + Fjob + reason + guardian + traveltime + studytime +
    failures + schoolsup + famsup + paid + activities + nursery +
    higher + internet + romantic + famrel + freetime + goout +
    Dalc + Walc + health + absences + G2 + G3) - G2 - G3
Model 2: G1 ~ sex + Fedu + studytime + failures + schoolsup + famsup +
    goout
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       355     263.08
2       387     302.02 -32  -38.936   0.1858
```

```r
# Model 3: GLM with Walc instead of goout
model3 <- glm(G1 ~ sex +
                Fedu +
                studytime +
                failures +
                schoolsup +
                famsup +
                Walc,
              data = students.math,
              family = poisson())
summary(model3)
```

```
Call:
glm(formula = G1 ~ sex + Fedu + studytime + failures + schoolsup +
    famsup + Walc, family = poisson(), data = students.math)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.31120    0.07204  32.083  < 2e-16 ***
sexM          0.07558    0.03296   2.293  0.02183 *
Fedu          0.04067    0.01477   2.753  0.00591 **
studytime     0.05231    0.01935   2.704  0.00685 **
failures     -0.14110    0.02488  -5.671 1.42e-08 ***
schoolsupyes -0.20115    0.04985  -4.035 5.46e-05 ***
famsupyes    -0.07447    0.03239  -2.299  0.02148 *
Walc         -0.02614    0.01274  -2.052  0.04016 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 402.47  on 394  degrees of freedom
Residual deviance: 304.07  on 387  degrees of freedom
AIC: 1977.2

Number of Fisher Scoring iterations: 4
```

```r
# Goodness-of-fit: AIC value
cat("AIC value of Model 3:", AIC(model3), "\n")
```

```
AIC value of Model 3: 1977.193
```

**Model 2**

- Significance of Covariates: All covariates are significant, with some even being very or highly significant.
- Goodness-of-Fit: The AIC value of Model 2 is smaller than the AIC value of Model 1, suggesting that Model 2 has a better fit. However, the difference is rather small.
- Analysis of Deviance Test: The p-value (0.1858) is greater than the conventional significance level (e.g., 0.05). This means that the reduction in deviance by adding the additional variables in Model 1 is not statistically significant. There is not enough evidence to suggest that the additional variables in Model 1 significantly improve the model fit compared to the simpler Model 2. Thus, Model 2, with fewer variables, is preferable.

**Model 3**

- Model 2 and Model 3 can be compared by the AIC value and the significance levels of the covariates.
- AIC Comparison: The AIC value of Model 3 is slightly higher than that of Model 2. However, the difference is only marginal.
- Since the significance levels of the covariates did not change and the AIC values are so close together, one could be indifferent between Model 2 and Model 3.