Homework 9 Question 2

Temitope B Adeniyi

2024-12-07

```r
# Load necessary library
library(MASS)

# Data input
quine <- read.csv("/Users/temitopeadeniyi/Downloads/quine.csv")

# Check mean and variance of Days
mean_days <- mean(quine$Days)
var_days <- var(quine$Days)

# Output results
cat("Mean of Days:", round(mean_days, 3), "\n")
```

```
## Mean of Days: 16.459
```

```r
cat("Variance of Days:", round(var_days, 3), "\n")
```

```
## Variance of Days: 264.167
```

```r
# Overdispersion check
if (var_days > mean_days) {
  cat("Overdispersion is present.\n")
} else {
  cat("No overdispersion detected.\n")
}
```

```
## Overdispersion is present.
```

```r
# Calculating by hand
# Step 1: Calculate Group Means
# Mean for Females (Sex = "F")
mu_F <- mean(quine$Days[quine$Sex == "F"])

# Mean for Males (Sex = "M")
mu_M <- mean(quine$Days[quine$Sex == "M"])

# Print the group means
cat("Mean for Females (mu_F):", round(mu_F, 3), "\n")
```

```
## Mean for Females (mu_F): 15.225
```

```r
cat("Mean for Males (mu_M):", round(mu_M, 3), "\n")
```

```
## Mean for Males (mu_M): 17.955
```

```r
# Step 2: Estimate alpha (Intercept) and beta (Coefficient)
alpha_hat <- log(mu_F)  # Log of the mean for Females
beta_hat <- log(mu_M / mu_F)  # Log of the ratio of the means

# Print the estimates
cat("Estimated alpha (log(mu_F)) by hand:", round(alpha_hat, 3), "\n")
```

```
## Estimated alpha (log(mu_F)) by hand: 2.723
```

```r
cat("Estimated beta (log(mu_M / mu_F)) by hand:", round(beta_hat, 3), "\n")
```

```
## Estimated beta (log(mu_M / mu_F)) by hand: 0.165
```

```r
# Step 3: Verify results
# Fit Poisson regression model
# Poisson regression using Sex as predictor
model_sex <- glm(Days ~ Sex, family = poisson(link = "log"), data = quine)

# Model summary
summary(model_sex)
```

```
## 
## Call:
## glm(formula = Days ~ Sex, family = poisson(link = "log"), data = quine)
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  2.72294    0.02865  95.030  < 2e-16 ***
## SexM         0.16490    0.04080   4.041 5.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 2073.5  on 145  degrees of freedom
## Residual deviance: 2057.2  on 144  degrees of freedom
## AIC: 2649.7
## 
## Number of Fisher Scoring iterations: 5
```

```r
# Extract coefficients
alpha <- coef(model_sex)[1]
beta <- coef(model_sex)[2]
exp_beta <- exp(beta)

# Print results
cat("Intercept (alpha) with R:", round(alpha, 3), "\n")
```

```
## Intercept (alpha) with R: 2.723
```

```
cat("Coefficient for Sex (Beta) with R:", round(beta, 3), "\n")

## Coefficient for Sex (Beta) with R: 0.165

cat("Ratio of means (e^beta):", round(exp_beta, 3), "\n")

## Ratio of means (e^beta): 1.179

#Quest 2(d)
# Full model with all predictors
model_full <- glm(Days ~ Eth + Sex + Age + Lrn, family = poisson(link =
"log"), data = quine)

# Perform stepwise selection
model_step <- step(model_full, direction = "both")

## Start:  AIC=2299.18
## Days ~ Eth + Sex + Age + Lrn
##
##          Df Deviance    AIC
## <none>        1696.7 2299.2
## - Sex    1    1711.1 2311.6
## - Lrn    1    1742.5 2343.0
## - Age    3    1865.0 2461.5
## - Eth    1    1863.5 2464.0

# Summary of the final model
summary(model_step)

##
## Call:
## glm(formula = Days ~ Eth + Sex + Age + Lrn, family = poisson(link =
"log"),
##     data = quine)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.71538    0.06468  41.980  < 2e-16 ***
## EthN        -0.53360    0.04188 -12.740  < 2e-16 ***
## SexM         0.16160    0.04253   3.799 0.000145 ***
## AgeF1       -0.33390    0.07009  -4.764 1.90e-06 ***
## AgeF2        0.25783    0.06242   4.131 3.62e-05 ***
## AgeF3        0.42769    0.06769   6.319 2.64e-10 ***
## LrnSL        0.34894    0.05204   6.705 2.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2073.5  on 145  degrees of freedom
## Residual deviance: 1696.7  on 139  degrees of freedom
## AIC: 2299.2
##
## Number of Fisher Scoring iterations: 5
```

```
#Question 2(f)
# Fit Poisson model with interaction between Age and Sex
model_interact <- glm(Days ~ Eth + Sex * Age + Lrn, family = poisson(link =
"log"), data = quine)

# Summarize the model
summary(model_interact)
```

```
##
## Call:
## glm(formula = Days ~ Eth + Sex * Age + Lrn, family = poisson(link =
"log"),
##     data = quine)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.03906    0.07648  39.735  < 2e-16 ***
## EthN        -0.50548    0.04195 -12.050  < 2e-16 ***
## SexM        -0.46064    0.10063  -4.578 4.70e-06 ***
## AgeF1       -0.56561    0.09181  -6.161 7.25e-10 ***
## AgeF2       -0.31158    0.09733  -3.201  0.00137 **
## AgeF3       -0.16576    0.09670  -1.714  0.08652 .
## LrnSL        0.46041    0.05508   8.359  < 2e-16 ***
## SexM:AgeF1  -0.11243    0.15077  -0.746  0.45584
## SexM:AgeF2   0.91468    0.12754   7.172 7.40e-13 ***
## SexM:AgeF3   1.11221    0.12851   8.655  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2073.5  on 145  degrees of freedom
## Residual deviance: 1559.1  on 136  degrees of freedom
## AIC: 2167.5
##
## Number of Fisher Scoring iterations: 5
```

```
# Pairwise comparisons for specific groups
library(emmeans)
```

```
## Welcome to emmeans.
## Caution: You lose important information if you filter this package's
results.
## See '? untidy'
```

```
# i. Compare male and female students with Age = F1
emmeans(model_interact, pairwise ~ Sex | Age, at = list(Age = "F1"))
```

```
## $emmeans
## Age = F1:
##  Sex emmean     SE  df asymp.LCL asymp.UCL
##  F     2.45 0.0512 Inf      2.35      2.55
##  M     1.88 0.1020 Inf      1.68      2.08
##
## Results are averaged over the levels of: Eth, Lrn
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
##
## $contrasts
## Age = F1:
##  contrast estimate    SE  df z.ratio p.value
##  F - M       0.573 0.112 Inf   5.096  <.0001
##
## Results are averaged over the levels of: Eth, Lrn
## Results are given on the log (not the response) scale.
```

```
# ii. Compare female students with Age = F0 and Age = F3
emmeans(model_interact, pairwise ~ Age | Sex, at = list(Sex = "F"))
```

```
## $emmeans
## Sex = F:
##  Age emmean     SE  df asymp.LCL asymp.UCL
##  F0    3.02 0.0743 Inf      2.87      3.16
##  F1    2.45 0.0512 Inf      2.35      2.55
##  F2    2.70 0.0586 Inf      2.59      2.82
##  F3    2.85 0.0674 Inf      2.72      2.98
##
## Results are averaged over the levels of: Eth, Lrn
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
##
## $contrasts
## Sex = F:
##  contrast estimate     SE  df z.ratio p.value
##  F0 - F1     0.566 0.0918 Inf   6.161  <.0001
##  F0 - F2     0.312 0.0973 Inf   3.201  0.0075
##  F0 - F3     0.166 0.0967 Inf   1.714  0.3162
##  F1 - F2    -0.254 0.0732 Inf  -3.469  0.0029
##  F1 - F3    -0.400 0.0888 Inf  -4.501  <.0001
##  F2 - F3    -0.146 0.0961 Inf  -1.517  0.4272
##
## Results are averaged over the levels of: Eth, Lrn
## Results are given on the log (not the response) scale.
## P value adjustment: tukey method for comparing a family of 4 estimates
```

```
# iii. Compare male students with Age = F1 and Age = F2
emmeans(model_interact, pairwise ~ Age | Sex, at = list(Sex = "M", Age =
c("F1", "F2")))
```

```
## $emmeans
## Sex = M:
##  Age emmean     SE  df asymp.LCL asymp.UCL
##  F1    1.88 0.1020 Inf      1.68      2.08
##  F2    3.16 0.0456 Inf      3.07      3.25
##
## Results are averaged over the levels of: Eth, Lrn
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
##
## $contrasts
## Sex = M:
##  contrast estimate    SE  df z.ratio p.value
##  F1 - F2     -1.28 0.112 Inf -11.396  <.0001
##
## Results are averaged over the levels of: Eth, Lrn
## Results are given on the log (not the response) scale.
```

```
# iv. Compare female students with Age = F1 and Age = F3
emmeans(model_interact, pairwise ~ Age | Sex, at = list(Sex = "F", Age =
c("F1", "F3")))
```

```
## $emmeans
## Sex = F:
##  Age emmean     SE  df asymp.LCL asymp.UCL
##  F1    2.45 0.0512 Inf      2.35      2.55
##  F3    2.85 0.0674 Inf      2.72      2.98
##
## Results are averaged over the levels of: Eth, Lrn
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
##
## $contrasts
## Sex = F:
##  contrast estimate     SE  df z.ratio p.value
##  F1 - F3      -0.4 0.0888 Inf  -4.501  <.0001
##
## Results are averaged over the levels of: Eth, Lrn
## Results are given on the log (not the response) scale.
```

```
# v. Compare students: Sex = male, Age = F2 and Sex = female, Age = F3
emmeans(model_interact, pairwise ~ Sex * Age, at = list(Sex = c("M", "F"),
Age = c("F2", "F3")))
```

```
## $emmeans
##  Sex Age emmean     SE  df asymp.LCL asymp.UCL
##  M   F2    3.16 0.0456 Inf      3.07      3.25
##  F   F2    2.70 0.0586 Inf      2.59      2.82
##  M   F3    3.50 0.0584 Inf      3.39      3.62
##  F   F3    2.85 0.0674 Inf      2.72      2.98
##
```

```
## Results are averaged over the levels of: Eth, Lrn
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
##
## $contrasts
##  contrast    estimate     SE  df z.ratio p.value
##  M F2 - F F2    0.454 0.0748 Inf   6.071  <.0001
##  M F2 - M F3   -0.343 0.0726 Inf  -4.728  <.0001
##  M F2 - F F3    0.308 0.0801 Inf   3.849  0.0007
##  F F2 - M F3   -0.797 0.0900 Inf  -8.855  <.0001
##  F F2 - F F3   -0.146 0.0961 Inf  -1.517  0.4272
##  M F3 - F F3    0.652 0.0799 Inf   8.154  <.0001
##
## Results are averaged over the levels of: Eth, Lrn
## Results are given on the log (not the response) scale.
## P value adjustment: tukey method for comparing a family of 4 estimates
```

```
# Load the MASS package for Negative Binomial regression
library(MASS)

# Fit the Poisson model (for comparison)
model_poisson <- glm(Days ~ Eth + Sex + Age + Lrn, family = poisson(link =
"log"), data = quine)

# Fit the Negative Binomial model with identified predictors
model_nb <- glm.nb(Days ~ Eth + Sex + Age + Lrn, data = quine)

# Summarize both models
cat("Poisson Model Summary:\n")
```

```
## Poisson Model Summary:
```

```
summary(model_poisson)
```

```
##
## Call:
## glm(formula = Days ~ Eth + Sex + Age + Lrn, family = poisson(link =
"log"),
##     data = quine)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.71538    0.06468  41.980  < 2e-16 ***
## EthN        -0.53360    0.04188 -12.740  < 2e-16 ***
## SexM         0.16160    0.04253   3.799 0.000145 ***
## AgeF1       -0.33390    0.07009  -4.764 1.90e-06 ***
## AgeF2        0.25783    0.06242   4.131 3.62e-05 ***
## AgeF3        0.42769    0.06769   6.319 2.64e-10 ***
## LrnSL        0.34894    0.05204   6.705 2.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 2073.5  on 145  degrees of freedom
## Residual deviance: 1696.7  on 139  degrees of freedom
## AIC: 2299.2
## 
## Number of Fisher Scoring iterations: 5
```

```
cat("\nNegative Binomial Model Summary:\n")
```

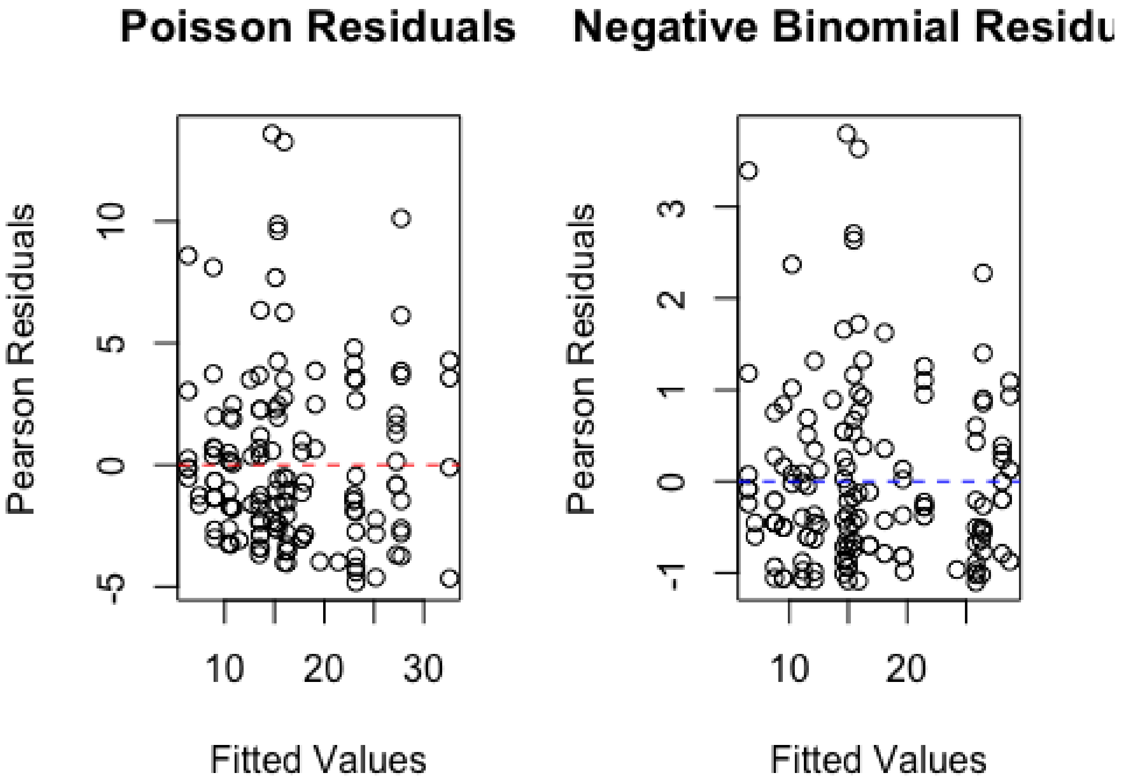
```
## 
## Negative Binomial Model Summary:
```

```
summary(model_nb)
```

```
## 
## Call:
## glm.nb(formula = Days ~ Eth + Sex + Age + Lrn, data = quine, 
##     init.theta = 1.274892646, link = log)
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  2.89458    0.22842  12.672  < 2e-16 ***
## EthN        -0.56937    0.15333  -3.713 0.000205 ***
## SexM         0.08232    0.15992   0.515 0.606710    
## AgeF1       -0.44843    0.23975  -1.870 0.061425 .  
## AgeF2        0.08808    0.23619   0.373 0.709211    
## AgeF3        0.35690    0.24832   1.437 0.150651    
## LrnSL        0.29211    0.18647   1.566 0.117236    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Negative Binomial(1.2749) family taken to be 1)
## 
##     Null deviance: 195.29  on 145  degrees of freedom
## Residual deviance: 167.95  on 139  degrees of freedom
## AIC: 1109.2
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##               Theta:  1.275 
##           Std. Err.:  0.161 
## 
##  2 x log-likelihood:  -1093.151
```

```
# Compare AIC values for model selection
cat("\nAIC for Poisson Model:", AIC(model_poisson), "\n")
```

```
##
## AIC for Poisson Model: 2299.184

cat("AIC for Negative Binomial Model:", AIC(model_nb), "\n")

## AIC for Negative Binomial Model: 1109.151

# Plot Residuals for Poisson and Negative Binomial Models
par(mfrow = c(1, 2))  # Set up a 1x2 plot grid

# Poisson residuals
plot(fitted(model_poisson), residuals(model_poisson, type = "pearson"),
     main = "Poisson Residuals", xlab = "Fitted Values", ylab = "Pearson
Residuals")
abline(h = 0, col = "red", lty = 2)

# Negative Binomial residuals
plot(fitted(model_nb), residuals(model_nb, type = "pearson"),
     main = "Negative Binomial Residuals", xlab = "Fitted Values", ylab =
"Pearson Residuals")
abline(h = 0, col = "blue", lty = 2)
```



```
summary(model_nb)
```

```
##
## Call:
## glm.nb(formula = Days ~ Eth + Sex + Age + Lrn, data = quine,
##     init.theta = 1.274892646, link = log)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.89458    0.22842  12.672  < 2e-16 ***
## EthN        -0.56937    0.15333  -3.713 0.000205 ***
## SexM         0.08232    0.15992   0.515 0.606710
## AgeF1       -0.44843    0.23975  -1.870 0.061425 .
## AgeF2        0.08808    0.23619   0.373 0.709211
## AgeF3        0.35690    0.24832   1.437 0.150651
## LrnSL        0.29211    0.18647   1.566 0.117236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.2749) family taken to be 1)
##
##     Null deviance: 195.29  on 145  degrees of freedom
## Residual deviance: 167.95  on 139  degrees of freedom
## AIC: 1109.2
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.275
##           Std. Err.:  0.161
##
##  2 x log-likelihood:  -1093.151
```