STAA 552: HW 6

YOUR NAME HERE

See Canvas Calendar for due date.
56 points total, 4 points per problem unless otherwise noted.
Add or delete code chunks as needed.
Content for Q1-Q8 is from section 09 or earlier.
Content for Q9-Q13 is from section 10 or earlier.

Toxicity (Q1 - Q8)

This data is from CDA. Rodent studies are commonly used to test and regulate substances posing potential danger to developing fetuses. This study administered an industrial solvent to pregnant mice. Each mouse was exposed to one of five concentration levels (0, 62.5, 125, 250, 500) for 10 days early in pregnancy. Later the fetuses were classified as dead, malformed or normal.

The data is available from Canvas as Toxicity.csv.

Q1

Start by summarizing the data by converting the counts to (row) proportions. Specifically, give the proportion dead, malformed and normal for each concentration.

```
##   concentration dead malformation normal
## 1           0.0   15            1    281
## 2          62.5   17            0    225
## 3         125.0   22            7    283
## 4         250.0   38           59    202
## 5         500.0  144          132      9
```

```
## # A tibble: 5 × 4
## # Rowwise:
##   concentration Prop_Dead Prop_Malformed Prop_Normal
##           <dbl>     <dbl>          <dbl>       <dbl>
## 1           0      0.0505        0.00337      0.946
## 2          62.5    0.0702        0            0.930
## 3         125      0.0705        0.0224       0.907
## 4         250      0.127         0.197        0.676
## 5         500      0.505         0.463        0.0316
```

Toxicity Nominal Logistic Regression (Q2 - Q5)

Fit a **baseline category logit** regression model to the data. This model treats the response as **nominal**.

Q2 (2 pts)

Fit the model using the provided code and show the coefficients table.

Notes:
(1) Dead = 1, Malformation = 2, Normal = 3 (baseline/reference).
(2) This code may generate warnings, but proceed with analysis.

```
## Warning: package 'VGAM' was built under R version 4.4.2

## Warning in vglm.fitter(x = x, y = y, w = w, offset = offset, Xm2 = Xm2, :
some
## quantities such as z, residuals, SEs may be inaccurate due to convergence
at a
## half-step

## Call:
## vglm(formula = cbind(dead, malformation, normal) ~ concentration,
##     family = multinomial, data = ToxData)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   -3.9694509  0.1909845  -20.78   <2e-16 ***
## (Intercept):2   -4.9527869  0.2493654  -19.86   <2e-16 ***
## concentration:1  0.0119089  0.0006958   17.11   <2e-16 ***
## concentration:2  0.0140096  0.0007867   17.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])
##
## Residual deviance: 57.5047 on 6 degrees of freedom
##
## Log-likelihood: -49.1804 on 6 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):1', '(Intercept):2'
##
##
## Reference group is level  3  of the response
```

Q3 (8 pts)

Calculate the two estimated conditional odds ratios from this model (dead vs normal, malformed vs normal) and provide a **detailed interpretation** for each.

Response

```
##   (Intercept):1   (Intercept):2 concentration:1 concentration:2 
##     0.018883798     0.007063696     1.011980103     1.014108184
```

- For each one unit increase in concentration, the odds of malformation increase by 1.2% (multiplicative effect: initial odds ration * 1.012))
- For each one unit increase in concentration, the odds of death increase by 1.4% (multiplicative effect: initial odds ration * 1.014)

Q4

Use a likelihood ratio test to test for an effect of concentration (versus the null model) and make a conclusion in context.

Response

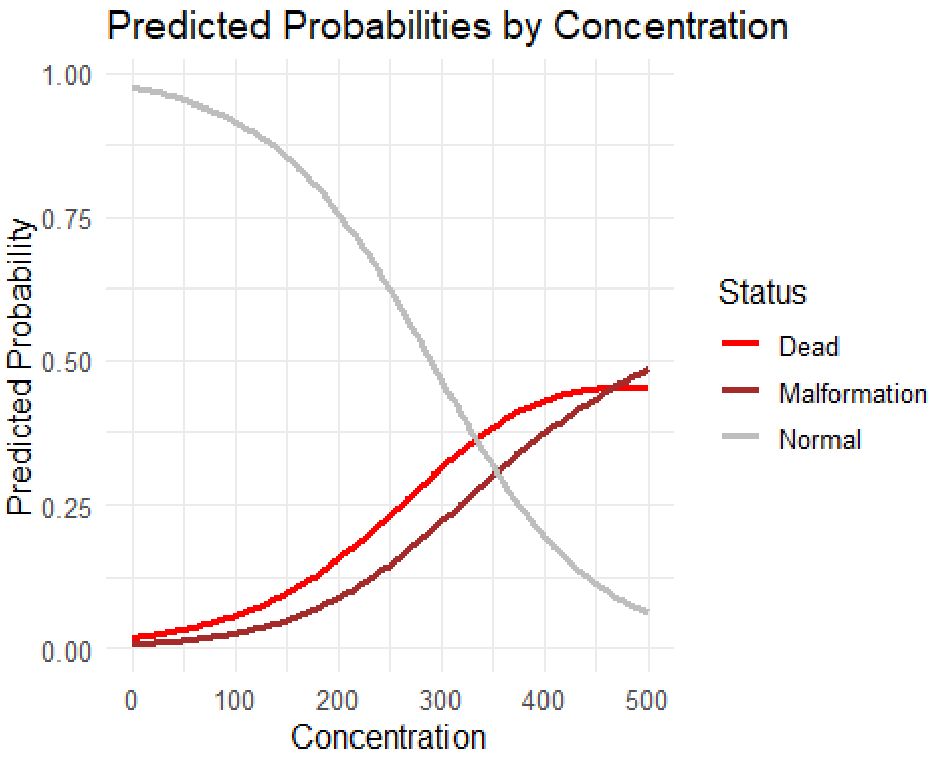
```
## Likelihood Ratio Statistic: 854.1856

## p-value: 3.280579e-186

## Likelihood ratio test
## 
## Model 1: cbind(dead, malformation, normal) ~ concentration
## Model 2: cbind(dead, malformation, normal) ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)    
## 1   6  -49.18                         
## 2   8 -476.27  2 854.19  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can conclude that our model is signifficantly better than the null model with a very very low p value *****

Q5 (6 pts)

Create a graph of the (3) fitted curves representing the predicted probabilities (dead, malformed and normal) over the range of concentration.

```
## 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.
```



Toxicity Ordinal Logistic Regression (Q6 - Q8)

Fit a **proportional odds** regression model to the data. This model treats the response as **ordinal**.

Q6 (2 pts)

Fit the model using the provided code and show the coefficients table.

Notes:
(1) Dead < Malformation < Normal.
(2) This code may generate warnings, but proceed with analysis.

```
## Call:
## vglm(formula = cbind(dead, malformation, normal) ~ concentration,
##     family = cumulative(parallel = TRUE), data = ToxData)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -4.5310552  0.1782478  -25.42   <2e-16 ***
## (Intercept):2 -3.1533473  0.1380590  -22.84   <2e-16 ***
## concentration  0.0096183  0.0004396   21.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 191.4714 on 7 degrees of freedom
##
## Log-likelihood: -116.1637 on 7 degrees of freedom
##
## Number of Fisher scoring iterations: 6
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2'
##
##
## Exponentiated coefficients:
## concentration
##      1.009665
```

Q7

Provide the estimated odds ratio corresponding to concentration and provide a **detailed interpretation**.

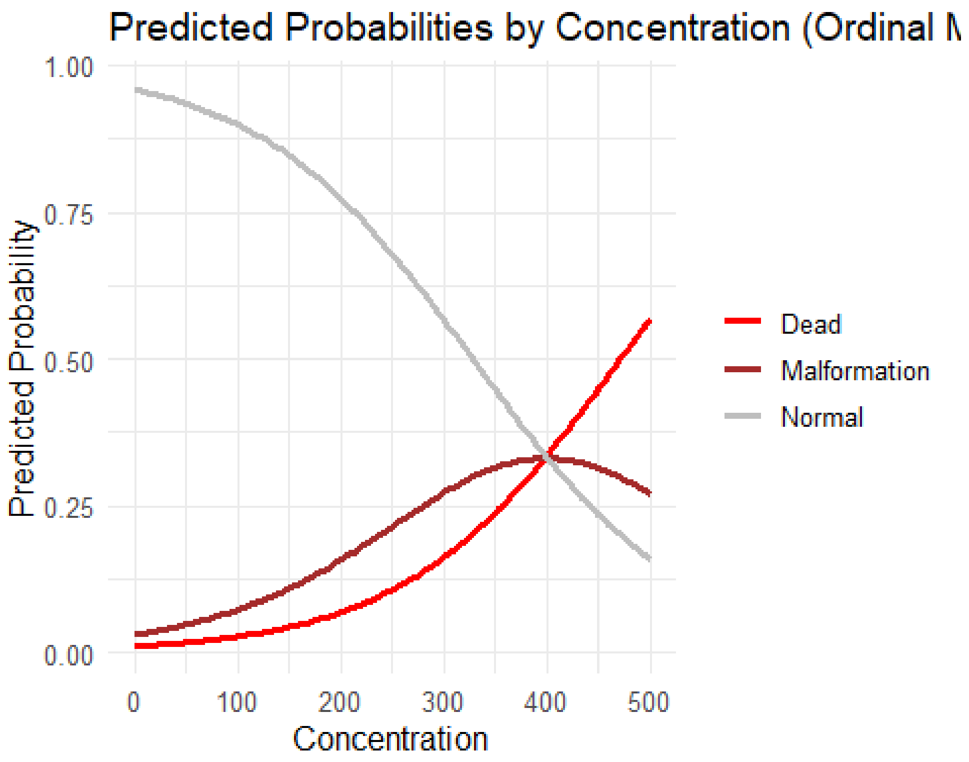
```
## (Intercept):1 (Intercept):2 concentration
##    0.01076931    0.04270893    1.00966471
```

Response

For each 1 unit increase in the concentration, the odds of being in the next higher category of normal > malformed > dead increase by .9% *****

Q8 (6 pts)

Create a graph of the (3) fitted curves representing the predicted probabilities (dead, malformed and normal) over the range of concentration.



Harbor Seals (Q9 - Q13)

Harbor seals in Alaska “haul out” onto landing sites to rest and warm themselves. While hauled out, they are relatively easy to count during aerial surveys (see Figure on Canvas).

Ecologists are interested in determining an “optimal” date in late summer or fall to conduct counts, meaning a date on which maximum numbers of hauled-out harbor seals could be counted. They consider a data set consisting of the count of SEALS by haul-out location (12 distinct sites, in the variable LOCNUMBER) and DATE, measured in days since August 15 (DATE = 0 for August 15, DATE = 1 for August 16, DATE = 2 for August 17, etc.) Historical count data across multiple years are in the data set Harbor_Seals.csv, available from Canvas.

Q9

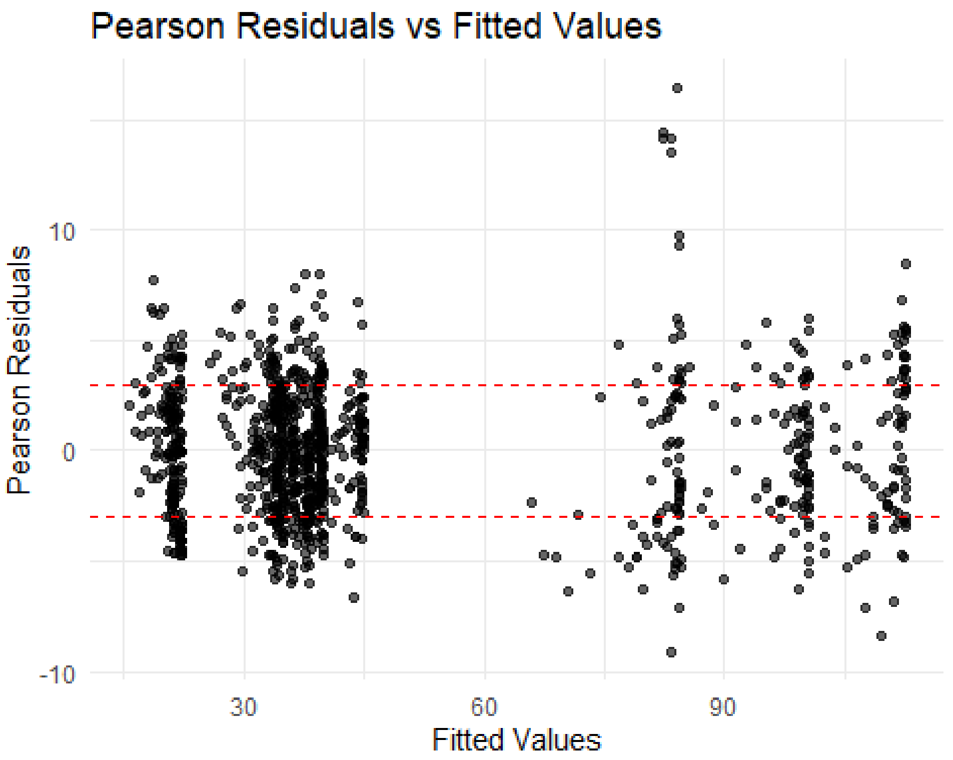
Fit a quasi-poisson model (with link = log) including LOCNUMBER (as.factor) and including DATA as a quadratic. This can be done using the poly(DATE, 2). Show the summary output and discuss whether over-dispersion is present in these data.

```
##
## Call:
## glm(formula = SEALS ~ factor(LOCNUMBER) + poly(DATE, 2), family =
quasipoisson(link = "log"),
##     data = Harbor_Seals)
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.76507    0.05040  74.698  < 2e-16 ***
## factor(LOCNUMBER)13 -0.70317    0.08597  -8.179 8.32e-16 ***
## factor(LOCNUMBER)14 -0.27281    0.07636  -3.573 0.000370 ***
## factor(LOCNUMBER)15 -0.11464    0.07353  -1.559 0.119268
## factor(LOCNUMBER)16  0.92075    0.05958  15.455  < 2e-16 ***
## factor(LOCNUMBER)17 -0.74979    0.08862  -8.461  < 2e-16 ***
## factor(LOCNUMBER)18  0.63314    0.06322  10.015  < 2e-16 ***
## factor(LOCNUMBER)19 -0.25181    0.07715  -3.264 0.001135 **
## factor(LOCNUMBER)20  0.80781    0.06080  13.285  < 2e-16 ***
## factor(LOCNUMBER)21 -0.21128    0.07586  -2.785 0.005450 **
## factor(LOCNUMBER)23 -0.12525    0.07296  -1.717 0.086359 .
## factor(LOCNUMBER)24 -0.16272    0.07434  -2.189 0.028831 *
## poly(DATE, 2)1      -1.74770    0.46602  -3.750 0.000186 ***
## poly(DATE, 2)2      -0.87839    0.46273  -1.898 0.057940 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 9.643431)
##
##     Null deviance: 25997  on 1049  degrees of freedom
## Residual deviance: 10544  on 1036  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

Comment Our Dispersion parameter is at 9.6. this is significantly high may indicate over dispersion in the data. *****

Q10

Further, construct a plot of Pearson residuals vs fitted values. Add horizontal references lines at +/-3. Does this diagnostic plot suggest that over-dispersion is present in these data? Briefly discuss. Note: The residual plot is the same whether we use poisson or quasi-poisson.



Comment There are way too many data points outside of the 3sd range. This indicates significant over dispersion as discussed above. *****

Q11 (2 pts)

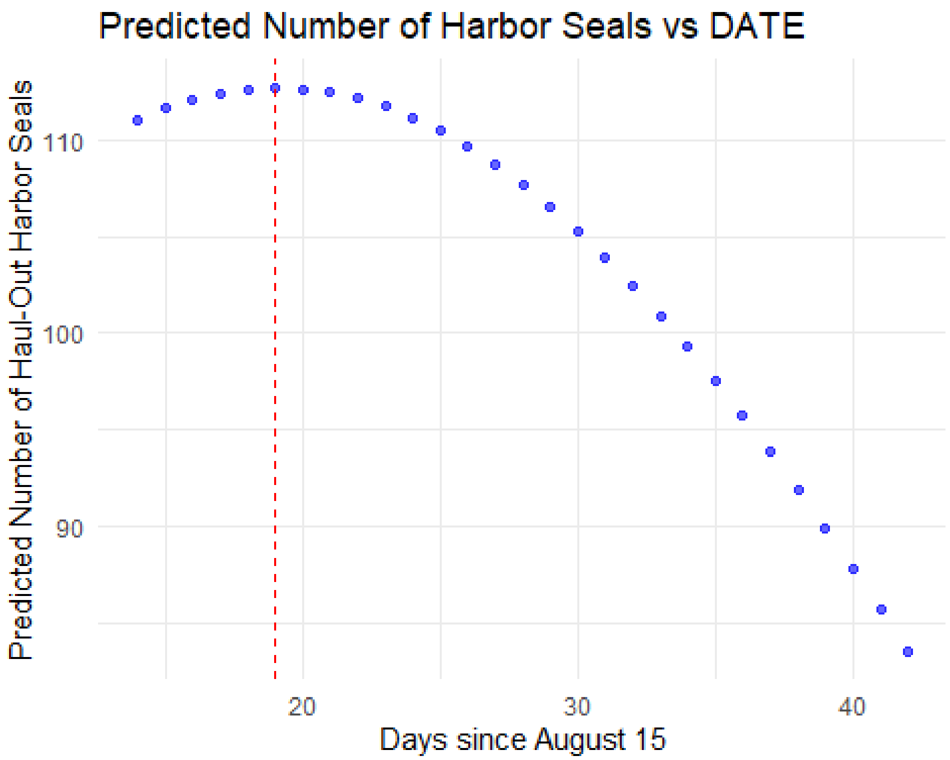
Use your fitted model to predict the number of hauled-out harbor seals at location 16, 33 days after August 15.

```
## Predicted number of hauled-out harbor seals at location 16 on day 33:
100.92
```

Q12 (6 pts)

Plot the predicted values versus DATE and address the ecologists' question about an optimum date for aerial surveys. What is the optimum date? Add a vertical line at that date. Note: For this question, you do not need to create smooth curves. Just plotting the predicted points is fine.

Optimum date is ?



```
## Optimum Date is: 19  Days after August 15th
```

Q13

Consider the plot from the previous question.

(a, 2pts) Explain the importance of the **quadratic** term for determining "optimum" date. In other words, what would happen if we had included DATE, but without the quadratic?

Response The quadratic term allows the possibility that the relationship between date and the number os seals is nonlinear. This is expected as these data tend to be cyclic. By having a quadratic term we get something that is logically maximizeable whereas a linear relationship maximizes either at 0 or at the max value. *****

(b, 2pts) How would the "optimum" date be affected if we included an **interaction** between LOCNUMBER and DATE?

Response This would allow the optimum date to vary depending on which location we are looking at. As a result, there may be multiple optimum dates (not if you take the global maximum, but for practical purposes).if you are looking on a per Location basis. This may actually be ideal as it is not reasonable for ecologists to photograph all locations on the same day. By allowing for different optimum dates per location you could photograph multiple different locations on its respective maximum day. *****

Appendix

```
#Retain this code chunk!!!
library(knitr)
library(tidyverse)
library(dplyr)
library(tidyr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
#Q1
ToxData <- read.csv("Data/Toxicity.csv")

print(ToxData)

proportion_data <- ToxData %>%
  rowwise() %>%
  mutate(Total = dead + malformation + normal) %>%
  mutate(
    Prop_Dead = dead / Total,
    Prop_Malformed = malformation / Total,
    Prop_Normal = normal / Total
  ) %>%
  select(concentration, Prop_Dead, Prop_Malformed, Prop_Normal)

# Display the proportions
print(proportion_data)
#Q2
library(VGAM)
ToxModel1 <- vglm(cbind(dead, malformation, normal) ~ concentration,
                  family = multinomial, data = ToxData)

summary_model1 <- summary(ToxModel1)
print(summary_model1)




#Q3
odds_ratios <- exp(coef(ToxModel1))

print(odds_ratios)
#Q4
ToxModel_null <- vglm(cbind(dead, malformation, normal) ~ 1,
                  family = multinomial, data = ToxData)
```

```r
logLik_null <- logLik(ToxModel_null)
logLik_full <- logLik(ToxModel1)

LR_stat <- 2*((logLik_full) - logLik_null)

p_value <- pchisq(LR_stat, df = 2, lower.tail = FALSE)


cat("Likelihood Ratio Statistic:", LR_stat, "\n")
cat("p-value:", p_value, "\n")

lrtest(ToxModel1,ToxModel_null)

#Q5
library(ggplot2)
library(reshape2)

new_conc <- data.frame(concentration = seq(min(ToxData$concentration),
                                          max(ToxData$concentration),
                                          length.out = 100))

pred_probs <- predict(ToxModel1, newdata = new_conc, type = "response")

plot_data <- cbind(new_conc, pred_probs)
plot_data_melt <- melt(plot_data, id.vars = "concentration",
                       variable.name = "Status",
                       value.name = "Probability")

# Plot using ggplot2
ggplot(plot_data_melt, aes(x = concentration, y = Probability, color =
Status)) +
  geom_line(size = 1.2) +
  labs(title = "Predicted Probabilities by Concentration",
       x = "Concentration",
       y = "Predicted Probability") +
  theme_minimal() +
  scale_color_manual(values = c("red", "brown", "grey"),
                     labels = c("Dead", "Malformation", "Normal")) +
  theme(text = element_text(size = 12))
#Q6
ToxModel2 <- vglm(cbind(dead, malformation, normal) ~ concentration,
                  family = cumulative(parallel=TRUE), data = ToxData)

summary2 = summary(ToxModel2)
print(summary2)



odds_ratios <- exp(coef(ToxModel2))
```

```r
print(odds_ratios)
#Q8

cumulative_probs <- predict(ToxModel2, newdata = new_conc, type = "response")

# Create the plot directly using ggplot2
ggplot() +
  geom_line(aes(x = new_conc$concentration, y = cumulative_probs[, 1], color
= "Dead"), size = 1.2) +
  geom_line(aes(x = new_conc$concentration, y = cumulative_probs[, 2], color
= "Malformation"), size = 1.2) +
  geom_line(aes(x = new_conc$concentration, y = cumulative_probs[, 3], color
= "Normal"), size = 1.2) +
  labs(title = "Predicted Probabilities by Concentration (Ordinal Model)",
       x = "Concentration",
       y = "Predicted Probability") +
  scale_color_manual(values = c("red", "brown", "grey"),
                     labels = c("Dead", "Malformation", "Normal")) +
  theme_minimal() +
  theme(legend.title = element_blank(),
        text = element_text(size = 12))



#Q9

# Read the data
Harbor_Seals <- read.csv("Data/Harbor_Seals.csv")

# Fit the quasi-Poisson model with LOCNUMBER as a factor and DATE as a
quadratic
model_quasi_pois <- glm(SEALS ~ factor(LOCNUMBER) + poly(DATE, 2),
                        family = quasipoisson(link = "log"),
                        data = Harbor_Seals)

# Display the summary of the model
summary(model_quasi_pois)
#Q10
ggplot() +
  geom_point(aes(x = fitted(model_quasi_pois), y =
residuals(model_quasi_pois, type = "pearson")),
             alpha = 0.6) +
  geom_hline(yintercept = c(-3, 3), linetype = "dashed", color = "red") +
  labs(title = "Pearson Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Pearson Residuals") +
  theme_minimal()
```

```
#Q11

new_data_Q11 <- data.frame(
  LOCNUMBER = factor(16, levels = unique(Harbor_Seals$LOCNUMBER)),
  DATE = 33
)

predicted_count <- predict(model_quasi_pois, newdata = new_data_Q11, type =
"response")

cat("Predicted number of hauled-out harbor seals at location 16 on day 33:",
round(predicted_count, 2))


#Q12
# Generate a sequence and predict date values
DATE_seq <- seq(min(Harbor_Seals$DATE), max(Harbor_Seals$DATE), by = 1)
predicted_counts <- predict(model_quasi_pois,
                            newdata = data.frame(LOCNUMBER = factor(16,
levels = unique(Harbor_Seals$LOCNUMBER)),
                                                 DATE = DATE_seq),
                            type = "response")

optimum_date <- DATE_seq[which.max(predicted_counts)]


# Plot the predicted counts vs DATE
ggplot() +
  geom_point(aes(x = DATE_seq, y = predicted_counts), color = "blue", alpha =
0.6) +
  geom_vline(xintercept = optimum_date, linetype = "dashed", color = "red") +
  labs(title = "Predicted Number of Harbor Seals vs DATE",
       x = "Days since August 15",
       y = "Predicted Number of Haul-Out Harbor Seals") +
  theme_minimal()

cat("Optimum Date is:", optimum_date, " Days after August 15th")
```