Generalized Linear Regression Homework 5

Brian Cervantes Alvarez
November 26, 2024

Problem 1

Part A

The Inverse Gaussian distribution is given by,

$$f(y) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right), \quad y > 0$$

we can expand it by,

$$-\frac{\lambda(y-\mu)^2}{2\mu^2 y} = -\frac{\lambda y}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2y}$$

Thus, we rewrite the PDF in the exponential family form,

$$f(y) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left(\theta_1 y + \theta_2 \frac{1}{y} - b(\theta)\right)$$

From this form, we can identify the parameters!

- **Canonical Parameters:** $\theta_1 = -\frac{\lambda}{2\mu^2}$, $\theta_2 = -\frac{\lambda}{2}$
- **Nuisance Parameter:** $\phi = 1$
- **Functions:**
  - $a(\phi) = 1$
  - $b(\theta) = -\frac{\lambda}{\mu}$
  - $c(y, \phi) = \sqrt{\frac{\lambda}{2\pi y^3}}$
- **Mean:** μ
- **Variance:** $\frac{\mu^3}{\lambda}$
- **Canonical Link Function:** $g(\mu) = \frac{1}{\mu}$

Part B

The Exponential distribution is defined as,

$$f(y) = \lambda e^{-\lambda y}, \quad y \geq 0$$

Rewriting the PDF in the exponential family form,

$$f(y) = \exp\left(-\lambda y + \log \lambda\right)$$

Again, now we can identify the parameters from this form,

- **Canonical Parameter:** $\theta = -\lambda$
- **Nuisance Parameter:** $\phi = 1$
- **Functions:**
  - $a(\phi) = 1$
  - $b(\theta) = 0$
  - $c(y, \phi) = \log \lambda$
- **Mean:** $\mu = \frac{1}{\lambda}$
- **Variance:** $\frac{1}{\lambda^2}$
- **Canonical Link Function:** $g(\mu) = \frac{1}{\mu}$

Problem 2

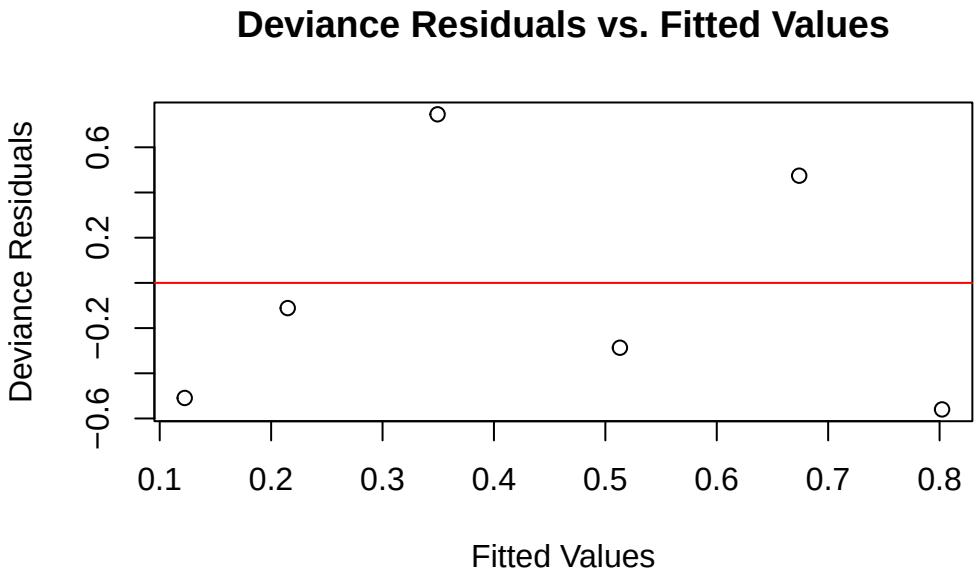
Part A

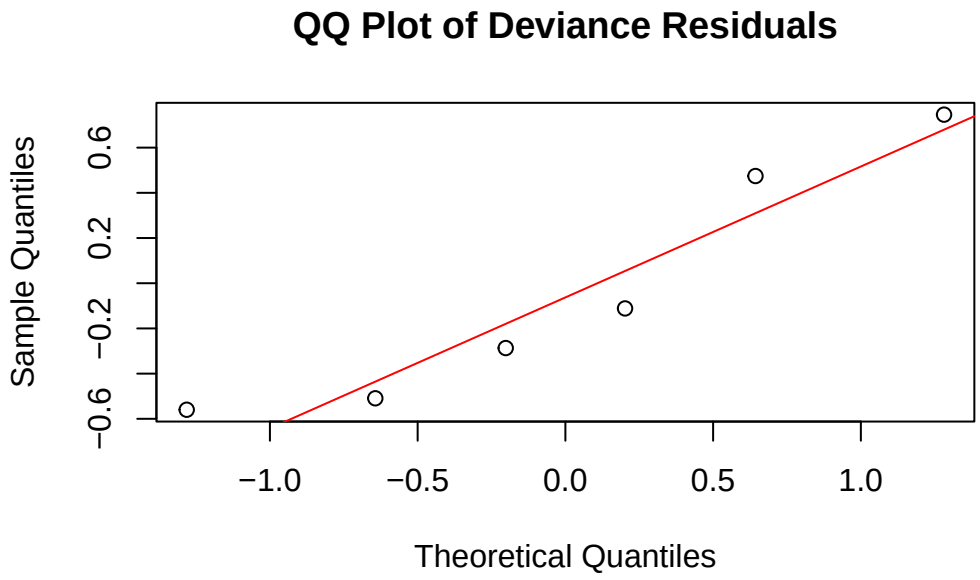
We can model the probability of death using logistic regression as follows,

$$\text{logit}(p) = \beta_0 + \beta_1 X$$

where $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$. After preparing the data, I fit the model using R,

Here, we extract and plot the deviance residuals to assess the model fit,





The QQ plot shows that most residuals align Ill with the theoretical quantiles, with only minor deviations at the tails, suggesting a reasonable fit to the model assumptions. The residuals vs. fitted values plot confirms this, with residuals randomly scattered around zero, indicating no clear evidence of model misspecification.

Part B

To evaluate the model fit, I calculated the deviance and its p-value,

- **Deviance:** 1.4491
- **Degrees of Freedom (Residual):** 4
- **p-value:** 0.8356

Since the p-value (0.8356) is greater than 0.05, I fail to reject the null hypothesis. This indicates that the model fits the data well.

Part C

If the observed variance exceeds the model's expected variance, then overdispersion is present. We can check this by finding the Pearson's Chi-Squared statistic,

- **Pearson's Chi-Squared (χ^2):** Approximately 1.4491
- **Degrees of Freedom (df):** 4

Then, we can compute the ratio χ^2/df,

$$\frac{\chi^2}{\text{df}} = \frac{1.4491}{4} \approx 0.3623$$

Since the ratio (≈ 0.3623) is less than 1, there is no evidence of overdispersion in the model.

Part D

Given that there is no overdispersion, the standard binomial model is adequate, and there is no need to fit a quasi-binomial model. However, for completeness, I fit a quasi-binomial model and performed a drop-in-deviance test,

Output:

```
Analysis of Deviance Table

Model 1: cbind(y, n - y) ~ 1
Model 2: cbind(y, n - y) ~ x
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         5     383.07
2         4       1.45  1   381.62 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Given the absence of overdispersion and the results of the drop-in-deviance test, we can suggest that the dose level x significantly influences mortality rates. Therefore, the full binomial model that includes x is statistically superior to the reduced model without x.

Part E

The estimated slope coefficient β_1 represents the change in the log-odds of death per unit increase in dose level x. A positive β_1 signifies that higher doses increase the probability of death. Then, I calculated the 95% confidence interval as follows,

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

Output

```
                 2.5 %     97.5 %
(Intercept) -2.9554809 -2.3432165
x            0.5985828  0.7519688
```

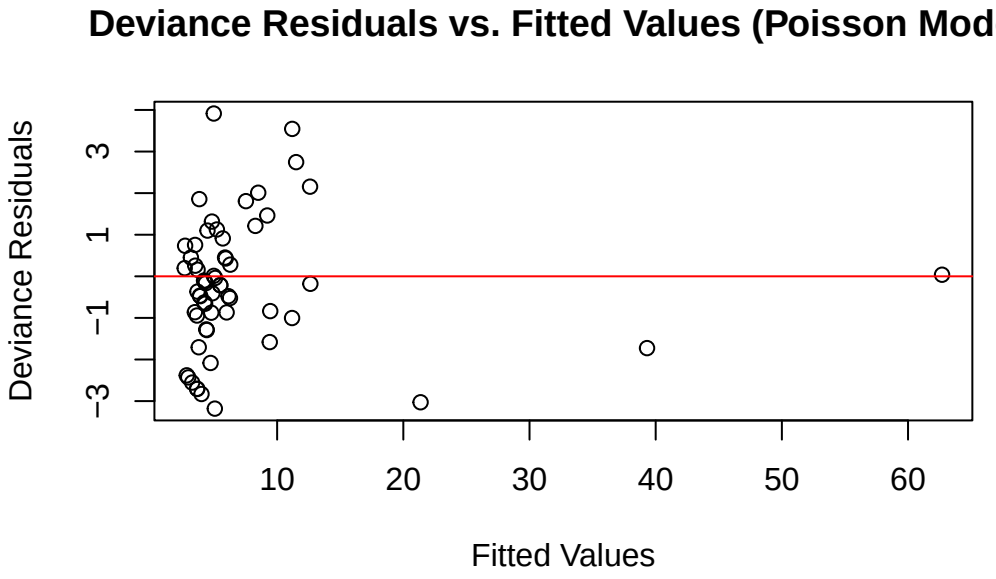
Since the entire confidence interval for β_1 is positive and does not include zero, we can suggest that there is a significant positive relationship betIen dose level and mortality. Specifically, for each unit increase in dose level x, the log-odds of death increase by approximately 0.60 to 0.75, indicating that higher doses are associated with higher mortality rates.

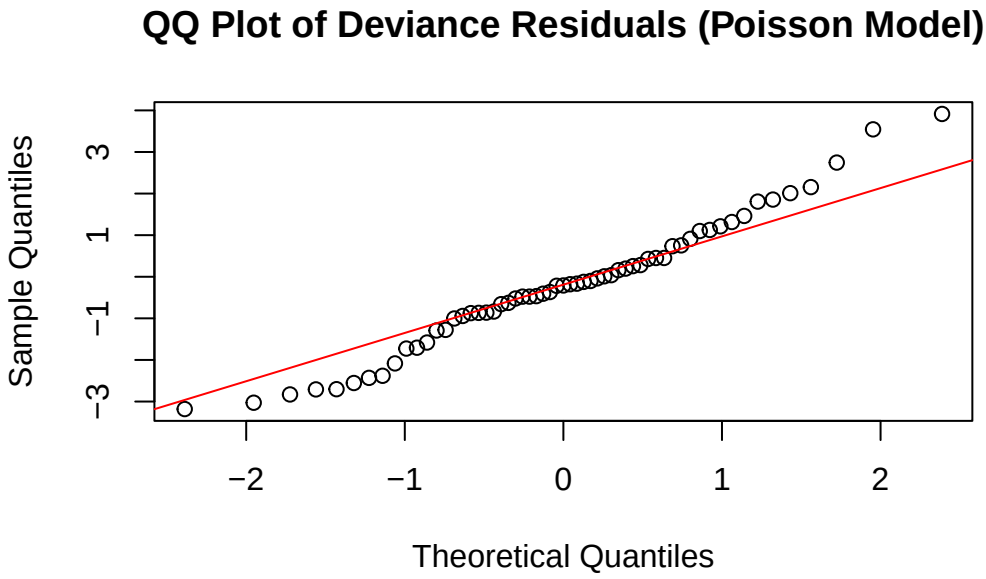
Problem 3

Part A

I analyzed the two-Iek seizure counts (`y4`) for epileptic patients using a Poisson regression model. The explanatory variables include the subject's age, treatment group (`trt`), and baseline seizure counts (`base`).

Next, I extracted the deviance residuals and plot them against the fitted values to assess the model fit. Additionally, I created a QQ plot to evaluate the normality of the residuals.





The Poisson model fits the data okay for most cases, but it struggles with extreme or large values. This might mean that the model needs adjustments, like accounting for extra variation.

Part B

To evaluate the adequacy of the Poisson regression model, we can perform the deviance goodness-of-fit test.

Output:

- **Deviance:** 144.5692
- **Degrees of Freedom (Residual):** 55
- **P-value:** 5.57×10^{-10}

Since the p-value is significantly less than 0.05, we reject the null hypothesis. This means that the Poisson model does not fit the data adequately.

Part C

I calculated Pearson's Chi-Squared statistic and its ratio to the degrees of freedom.

```
[1] "Pearson's Chi-Squared:133.585012943093"
```

```
[1] "Ratio (Chi2/df):2.42881841714714"
```

Output:

```
[1] "Pearson's Chi-Squared:133.585012943093"
[1] "Ratio (Chi2/df):2.42881841714714"
```

$$\frac{\chi^2}{\text{df}} = \frac{133.5850}{55} \approx 2.4288$$

There is evidence of overdispersion in the Poisson regression model, indicating that the variance in the data exceeds what the model expects. This suggests that the Poisson model may be inappropriate, and alternative models that account for overdispersion should be considered.

Part D

I fit a **quasi-Poisson** model, which accounts for overdispersion by introducing a dispersion parameter. Lastly, I performed a drop-in-deviance test to assess the significance of the treatment variable (`trt`).

Output:

```
Analysis of Deviance Table

Model 1: y4 ~ age + base
Model 2: y4 ~ age + trt + base
  Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
1        56     152.56
2        55     144.57  1   7.9918 3.2904 0.07514 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Residual Degrees of Freedom:** Reduced model has 56, full model has 55.
- **Residual Deviance:** Reduced model = 152.56, Full model = 144.57.
- **Deviance Difference:** $152.56 - 144.57 = 7.9918$
- **Degrees of Freedom for the Test:** 1
- **F-statistic:** 3.2904
- **P-value:** 0.07514

Even though the data shows extra variation that requires using a quasi-Poisson model, the treatment type doesn't play an important role in the results. This means that after considering age and the number of seizures at the start, the type of treatment (`trt`) doesn't have a meaningful effect on the number of seizures (`y4`) during the two-Iek period.