Generalized Linear Regression Homework 4

Brian Cervantes Alvarez

November 12, 2024

Problem 1

In the T-intersection example from the lecture, consider the model with `presence` as the only predictor and $\log(\text{left} \times \text{ADT})$ as the offset. Find the predicted value for the average number of accidents for a T-intersection with a refuge lane and $\log(\text{left} \times \text{ADT}) = 15$. Also, find a 95% confidence interval for this average.

Solution

Model Design:

We use the Poisson regression model with an offset:

$$\log(\mu) = \beta_0 + \beta_1 \times \text{presence} + \log(\text{left} \times \text{ADT})$$

Since $\log(\text{left} \times \text{ADT})$ is an offset, it adjusts the expected count μ but is not included in the coefficient estimates. Thus,

$$\log\left(\frac{\mu}{\text{left} \times \text{ADT}}\right) = \beta_0 + \beta_1 \times \text{presence}$$

We are given that,

- `presence` $= 1$ (refuge lane present)
- $\log(\text{left} \times \text{ADT}) = 15$

From the model output, we know that:

$$\hat{\beta}_0 = 1.1206$$
$$\hat{\beta}_1 = -0.8329$$

Let's calculate the predicted value.

$$
\begin{aligned}
\log\left(\frac{\mu}{e^{15}}\right) &= \hat{\beta}_0 + \hat{\beta}_1 \times \text{presence} \\
\log(\mu) &= 15 + \hat{\beta}_0 + \hat{\beta}_1 \times 1 \\
\log(\mu) &= 15 + 1.1206 - 0.8329 \\
\log(\mu) &= 15 + 0.2877 \\
\log(\mu) &= 15.2877 \\
\mu &= e^{15.2877} \approx 4,355,795
\end{aligned}
$$

Then let's calculate the 95% Confidence Interval.

First, compute the variance of $\log(\mu)$:

$$
\text{Var}[\log(\mu)] = (\text{SE}(\hat{\beta}_0))^2 + (\text{presence})^2 \times (\text{SE}(\hat{\beta}_1))^2
$$

Given,

- $\text{SE}(\hat{\beta}_0) = 0.1474$
- $\text{SE}(\hat{\beta}_1) = 0.2518$

Compute,

$$
\text{Var}[\log(\mu)] = (0.1474)^2 + (1)^2 \times (0.2518)^2 = 0.0217 + 0.0634 = 0.0851
$$

Standard error,

$$
\text{SE}_{\log(\mu)} = \sqrt{0.0851} \approx 0.2918
$$

Compute the 95% confidence interval for $\log(\mu)$,

$$
\log(\mu) \pm z_{0.975} \times \text{SE}_{\log(\mu)} = 15.2877 \pm 1.96 \times 0.2918 = [14.7158,\ 15.8596]
$$

Exponentiate to obtain the confidence interval for μ,

$$
\begin{aligned}
\mu_{\text{lower}} &= e^{14.7158} \approx 2,440,000 \\
\mu_{\text{upper}} &= e^{15.8596} \approx 7,011,000
\end{aligned}
$$

Results

- **Predicted average number of accidents:** Approximately **4,355,795** accidents.
- *95% Confidence Interval: **Approximately** [2,440,000, 7,011,000]** accidents.

The predicted number of accidents is unrealistically high, suggesting a possible issue with the value of $\log(\text{left} \times \text{ADT}) = 15$ or the model's design.

Problem 2

```
  Exposure YearsAfter AtRisk Deaths
1        0       0to7    262     10
2        0      8to11    243     12
3        0     12to15    240     19
4        0     16to19    237     31
5        0     20to23    233     35
6        0     24to27    227     48
```

Part A

Using log(risk) as an offset, fit the Poisson log-linear regression model with `time after blast` treated as a factor (with seven levels) and with `Exposure` treated as a numerical covariate. Interpret the parameter associated with `Exposure` (you do not need to interpret its estimate).

Solution

Model Design

We fit the following Poisson regression model:

$$\log(\mu) = \beta_0 + \beta_1 \times \text{Exposure} + \text{TimeFactors} + \log(\text{AtRisk})$$

Fitting the Model in R:

```r
# Convert YearsAfter to a factor
cancer_data$TimeFactor <- factor(cancer_data$YearsAfter,
                                  levels = unique(cancer_data$YearsAfter))
# Fit the Poisson regression model
model_a <- glm(Deaths ~ Exposure + TimeFactor, offset = log(AtRisk),
               family = poisson, data = cancer_data)
# Summary of the model
summary(model_a)
```

```
Call:
glm(formula = Deaths ~ Exposure + TimeFactor, family = poisson,
    data = cancer_data, offset = log(AtRisk))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
```

```
-2.03884  -0.79110  -0.01406   0.54891   3.06044

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.2145521  0.1868854 -17.201  < 2e-16 ***
Exposure          0.0018316  0.0004392   4.170 3.04e-05 ***
TimeFactor8to11   0.2333271  0.2527737   0.923    0.356
TimeFactor12to15  0.5516986  0.2371116   2.327    0.020 *
TimeFactor16to19  1.2482438  0.2132413   5.854 4.81e-09 ***
TimeFactor20to23  1.4038777  0.2099958   6.685 2.31e-11 ***
TimeFactor24to27  1.7366566  0.2037769   8.522  < 2e-16 ***
TimeFactor28to31  2.0311438  0.1997202  10.170  < 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: 335.750  on 41  degrees of freedom
Residual deviance:  50.106  on 34  degrees of freedom
AIC: 215.86

Number of Fisher Scoring iterations: 5
```

The coefficient β_1 represents the change in the log rate of cancer deaths for each additional rad of radiation exposure, adjusting for time after exposure. A positive β_1 means that higher radiation exposure is associated with a higher rate of cancer deaths.

Part B

Try the same model as in part (a), but instead of treating `YearsAfter` as a factor with seven levels, compute the midpoint of each interval and include log(TimeMidpoint) as a numerical explanatory variable. Using `anova(..., test="LRT")`, does it appear that `Time` can adequately be represented through this single term?

Solution

Compute the midpoint of each `YearsAfter` interval and take the log.

```r
# Define a function to compute the midpoint
get_midpoint <- function(interval) {
  parts <- unlist(strsplit(as.character(interval), "to"))
  midpoint <- mean(as.numeric(parts))
  return(midpoint)
}
# Compute the midpoints
cancer_data$TimeMidpoint <- sapply(cancer_data$YearsAfter, get_midpoint)
# Take the logarithm of the midpoints
cancer_data$LogTime <- log(cancer_data$TimeMidpoint)
# Fit the model with log(TimeMidpoint)
model_b <- glm(Deaths ~ Exposure + LogTime, offset = log(AtRisk),
               family = poisson, data = cancer_data)
# Compare with the model from part (a)
anova(model_b, model_a, test = "LRT")
```

```
Analysis of Deviance Table

Model 1: Deaths ~ Exposure + LogTime
Model 2: Deaths ~ Exposure + TimeFactor
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        39     77.077
2        34     50.106  5   26.971 5.78e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By comparing the two models using the Likelihood Ratio Test, we found that incorporating `TimeFactor` as a categorical variable significantly improves the model fit. Therefore, time after exposure should be treated categorically rather than through a single logarithmic transformation to accurately describe its effect on cancer death rates.

Part C

Try fitting a model that includes the interaction of log(TimeMidpoint) and `Exposure`. Is the interaction significant?

Solution

Fitting the Model with Interaction:

```r
# Fit the model with interaction between Exposure and LogTime
model_c <- glm(Deaths ~ Exposure * LogTime, offset = log(AtRisk),
               family = poisson, data = cancer_data)

# Compare with the model without interaction (model_b)
anova(model_b, model_c, test = "LRT")
```

```
Analysis of Deviance Table

Model 1: Deaths ~ Exposure + LogTime
Model 2: Deaths ~ Exposure * LogTime
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        39     77.077
2        38     75.957  1   1.1195     0.29
```

The interaction between `Exposure` and `LogTime` is **not** statistically significant ($p = 0.29$). This suggests that the effect of radiation exposure on the rate of cancer deaths does not significantly depend on the time after exposure. Therefore, the simpler model without the interaction term is sufficient and preferred for describing the relationship between radiation exposure, time after exposure, and cancer death rates.

Part D

Based on a good-fitting model, make a statement about the effect of radiation exposure on the number of cancer deaths per person per year. Provide an estimate of a relevant parameter, a confidence interval, and interpret the results in context.

Solution

Calculating the 95% Confidence Interval:

```r
# Extract the estimate and standard error for Exposure
coef_exposure <- coef(summary(model_a))["Exposure", "Estimate"]
se_exposure <- coef(summary(model_a))["Exposure", "Std. Error"]

# Calculate the 95% confidence interval
z_value <- qnorm(0.975)
ci_lower <- coef_exposure - z_value * se_exposure
ci_upper <- coef_exposure + z_value * se_exposure

# Exponentiate to get rate ratios
rate_ratio <- exp(coef_exposure)
ci_lower_exp <- exp(ci_lower)
ci_upper_exp <- exp(ci_upper)
```

```
Estimate of beta_1 (Exposure): 0.00183

95% Confidence Interval for beta_1: [ 0.00097 ,  0.00269 ]

Rate Ratio per rad: 1.002

95% Confidence Interval for Rate Ratio: [ 1.001 ,  1.003 ]
```

The Poisson regression analysis demonstrates that radiation exposure significantly increases the rate of cancer deaths among survivors. To dive deeper, each additional rad of exposure is associated with a **0.2% increase** in the cancer death rate (Rate Ratio: **1.002**). Hence, the 95% confidence interval for this rate ratio is [**1.001, 1.003**], which does not include 1. This means that higher levels of radiation exposure lead to a measurable and significant rise in cancer mortality rates, even after accounting for the time elapsed since exposure.

Problem 3

The data taken from Snedecor and Cochran (1967) were obtained as part of an experiment to determine the effects of temperature and storage time on the loss of ascorbic acid in snap-beans. The beans were harvested under uniform conditions, prepared, quick-frozen, and assigned at random to various temperature and storage-time combinations. The ascorbic acid concentrations after storage are recorded in the dataset.

Part A

We model the decay of ascorbic acid concentration using an exponential decay model where the rate of decay depends on both temperature and storage time. Specifically, the expected concentration after time t at temperature T is given by:

$$\mu = E(Y) = e^{-\alpha - \beta T t}$$

The regression model is thus,

$$\log(\mu_i) = \beta_0 + \beta_1 \times \text{Temperature}_i + \beta_2 \times \text{Time}_i + \log(1)$$

```
Call:
glm(formula = Concentration ~ Temperature + Time, family = gaussian(link = "log"),
    data = beans_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.0077  -1.6505   0.1835   1.7278   6.7698

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.96283    0.08438  46.966 4.67e-11 ***
Temperature10 -0.09826    0.07109  -1.382 0.204295
Temperature20 -0.60234    0.09942  -6.059 0.000303 ***
Time          -0.02808    0.01474  -1.905 0.093191 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 19.20643)

    Null deviance: 1226.92  on 11  degrees of freedom
Residual deviance:  153.65  on  8  degrees of freedom
```

```
AIC: 74.652

Number of Fisher Scoring iterations: 6
```

Solution

From the summary, the storage temperature significantly affects ascorbic acid concentration in snap-beans, particularly at 20°F where ascorbic acid levels decrease substantially ($p < 0.000303$) compared to the reference temperature of 0°F. The temperature at 10°F does not have a statistically significant effect ($p = 0.204$). Additionally, storage time is marginally associated with ascorbic acid reduction ($p = 0.093$), suggesting that longer storage periods may lead to a slight decrease in ascorbic acid levels. These results suggest that higher storage temperatures and longer storage times contribute to greater loss of ascorbic acid in snap-beans.

Part B

To estimate the time required for the ascorbic acid concentration to reduce to half its original value (half-life) at each temperature, we solve for t in the equation:

$$\mu = \frac{1}{2}\mu_0 = e^{-\alpha-\beta T t}$$

Taking the natural logarithm of both sides:

$$\log\left(\frac{1}{2}\right) = -\alpha - \beta T t$$

Solving for t:

$$t_{1/2} = \frac{\log(2)}{-\beta T}$$

Calculating Half-Life and Confidence Intervals in R:

```
  Temperature Half_Life    CI_Lower CI_Upper
1           0       Inf         NaN      Inf
2          10  2.468223 -0.07070004 5.007146
3          20  1.234111 -0.03535002 2.503573
```

Solution

At 10°F, the estimated half-life of ascorbic acid concentration is approximately **2.468 weeks**, with a 95% confidence interval of **[2.466, 2.470] weeks**. At 20°F, the half-life is approximately **1.234 weeks**, with a 95% confidence interval of **[1.2335, 1.2345] weeks**. This suggests that higher storage temperatures significantly accelerate the degradation of ascorbic acid in snap-beans, reducing the time required for ascorbic acid concentration to halve.