Beyond MLR Lab 7: Poisson regression models

Case study: Asthma exacerbations

Asthma exacerbations represent sudden worsening of asthma symptoms, often requiring urgent medical attention. Understanding the factors that increase the frequency of exacerbations can aid in risk assessment and prevention strategies. This case study focuses on analyzing individual-level data to explore how demographic and lifestyle factors contribute to asthma exacerbation rates.

Dataset

- The dataset `asthma_data.csv` contains information for 500 individuals.
- The columns in the dataset are:
  - `ID`: Unique identifier for each individual.
  - `Age`: Age of the individual in years.
  - `Smoking_Status`: Smoking status (1 = smoker, 0 = non-smoker).
  - `Exacerbations`: Number of asthma exacerbations experienced in one year.
- The dataset is available for download from Brightspace.

Exploratory Data Analysis

```
# Load required packages
library(ggplot2)
library(dplyr)
```

We begin by loading the dataset and examining its structure:

```r
# Load the dataset
asthma_data <- read.csv("asthma_data.csv")

# Convert Smoking_Status to a factor
asthma_data$Smoking_Status <- factor(asthma_data$Smoking_Status, labels = c("Non-Smoker", "Sm

# Inspect the structure of the dataset
str(asthma_data)
```

Next, we calculate summary statistics and create a bar chart to visualize the distribution of asthma exacerbations in the population:

```r
# Summary statistics
asthma_data %>%
  summarize(
    Mean_Exacerbations = mean(Exacerbations),
    Variance_Exacerbations = var(Exacerbations)
  )

# Bar chart of exacerbation counts
ggplot(asthma_data, aes(x = Exacerbations)) +
  geom_bar(fill = "skyblue", color = "black", alpha = 0.7) +
  labs(
    title = "Distribution of Asthma Exacerbations",
    x = "Number of Exacerbations",
    y = "Count of Individuals"
  ) +
  theme_minimal()
```

Question

Based on the bar chart and the summary statistics, describe the distribution of asthma exacerbations in the population.

Exercise

Create facetted bar charts showing the distribution of asthma exacerbations, stratified by:

- **Smoking Status** (facet by smoking group: Non-Smoker vs. Smoker).
- **Age Groups**.

Hint: Use the `cut()` function to create age categories. For example:

```r
asthma_data <- asthma_data %>%
  mutate(Age_Group = cut(Age, breaks = c(-Inf, 40, 60, Inf), labels = c("<40", "40-60", ">6
```

Exercise

Calculate the mean and variance of the number of exacerbations stratified by smoking status and age groups.

Question

What do the facetted bar charts reveal about differences in the distribution of exacerbations across smoking and age groups?

Question

Are the mean and variance of exacerbations consistent across subgroups? What might this imply about the data or the Poisson model's assumptions?

Fitting a Poisson Regression Model

The syntax in the code chunk below fits a Poisson regression model to investigate the relationship between asthma exacerbations and age and smoking status:

```r
# Fit the Poisson regression model
poisson_model <- glm(
  Exacerbations ~ Age + Smoking_Status,
  family = poisson(link = "log"),
  data = asthma_data
)

# View the model summary
summary(poisson_model)
```

Explanation of the code:

- `glm()` function: Fits a generalized linear model (GLM) to the data.
- `Exacerbations ~ Age + Smoking_Status`: Specifies the outcome (`Exacerbations`) and predictors (`Age` and `Smoking_Status`).
- `family = poisson(link = "log")`: Indicates that the model uses the Poisson distribution with a log link function.

Question

What does the intercept of the model represent in this context?

Question

Interpret the coefficient for Smoking_Status. How does smoking affect the expected number of asthma exacerbations?

Question

Interpret the coefficient for Age. How does an increase in age affect the expected number of asthma exacerbations?

Evaluating Model Fit: Pearson Residuals and Goodness-of-Fit

To evaluate the model fit, we calculate the **Pearson residuals**. These residuals measure the standardized difference between observed and predicted counts:

- Formula: $r_i = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$, where:
  - Y_i: Observed counts
  - $\hat{\lambda}_i$: Predicted counts

```
# Calculate Pearson residuals
asthma_data$Pearson_Residuals <- residuals(poisson_model, type = "pearson")
```

Next, we assess the goodness-of-fit using the **chi-square test**:

- Test statistic: $\chi^2 = \sum r_i^2 = \sum \frac{(Y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}$
- Degrees of freedom: $\text{df} = n - p$, where:
  - n: Number of observations.
  - p: Number of model parameters (including intercept).

```
# Calculate chi-square test statistic
chisq_test <- sum(asthma_data$Pearson_Residuals^2)

# Degrees of freedom
df <- nrow(asthma_data) - length(coef(poisson_model))
```

```r
# p-value
p_value <- 1 - pchisq(chisq_test, df)

# Display results
list(Chi_Square_Statistic = chisq_test, Degrees_of_Freedom = df, P_Value = p_value)
```

Question

What does the chi-square test tell us about the goodness-of-fit of the Poisson regression model?