**EPID 5314 - HW 3**

**Part 1:**

Chart, line chart

Description automatically generated

The Poisson regression model assumes that the variance function (V(µ)) is equal to the mean (µ); in other words, variance of the response variable is dependent on the mean. The full relationship is described by the equation below:

V(µ) = Φ\*µ, where Φ = 1

However, a phenomenon termed “overdispersion” may occur if the above equation does not hold for the data fitted to the Poisson model. More specifically, overdispersion occurs when the observed variance in a dataset is greater than the variance predicted by the Poisson model.

The probability plot (above) can be used to assess the appropriateness of the Poisson model for fitting our data. This graph is a plot of the SAS generated Pearson residuals against the percentiles for the normal distribution. Additionally, the plot shows whether or not the Pearson residuals are normally distributed, which occurs when the plotted residuals approximate the normal line (in blue). Since the plotted residuals above significantly deviate from linearity in the plot above, we can conclude that the Poisson model is not a good fit for this data. This *may* be the result of overdispersion.

**Part 2:**

A picture containing graphical user interface

Description automatically generated

The reported OR for the effect of ever having smoked a cigarette in 8th grade to having a larger number of friends smoke in 12th grade was 0.4277 with a 95% CI of 0.3587 to 0.5099. According to these findings the odds of having friends in the 12th grade who smoke changes by 0.4277 (decreases) for a 1 unit increase in the predictor variable (ever smoked in the 8th grade). This change, in terms of 95% CI, is between 0.3587 and 0.5099 for a 1 unit increase in the predictor variable. Since the null value (1) is not included in this 95% CI, the OR is significant.

Alternatively, we could interpret this as the odds of having more friends in the 12th grade who smoke cigarettes is lower for those in the “smoked a cigarette in 8th grade” (cigever4 = 1) category as compared to those in the “never smoked a cigarette in 8th grade” (cigever4 = 0) category.

Graphical user interface, text, application

Description automatically generated

To test the proportional odds assumption, I used the Score Test generated by proc logistic. For this test, an insignificant p-value (reported above) means that the odds ratios can be interpreted as constant across all levels of the outcome variable. This means that the assumption holds when the reported p-value is greater than α for your given level of confidence. According to this interpretation, the proportional odds assumption does hold for this data since the Score Test is non-significant at the 95% level of confidence (p = 0.6337 > 0.05).

**SAS Code**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Course: Data Analysis - EPID 5314 \*

\* Assignment: Homework 3 \*

\* Due Date: 9/15/2021 \*

\* Programmer(s): Jessie Ausman \*

\* Program Name: HW3 \*

\* Save Program/Log/Output: C:\Users\jessa\Desktop\EPID 5314\Homework3 \*

\* Save Data Files: C:\Users\jessa\Desktop\EPID 5314\PNC Data File\PNC \*

\* Datasets \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* PART 0 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Read in Datasets

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

libname bios "C:\Users\jessa\Desktop\EPID 5314\PNC Data File\PNC Datasets";

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Create merged dataset

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**data** pnc05;

set bios.pnc05;

**run**;

**data** pnc09;

set bios.pnc09;

**run**;

**data** lab3;

merge pnc05 (in=a) pnc09 (in=b);

by ID;

if a;

if b;

**run**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Part 1 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Generate Pearson's regression plots

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**proc** **freq** data=lab3;

table cigever4\*fritob5;

**run**;

**proc** **genmod** data=lab3;

model fritob5 = cigever4/ dist=p;

output out=pout reschi=rs;

**run**;

**proc** **univariate** data=pout noprint;

var rs;

probplot rs/ normal (mu=est sigma=est);

**run**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Part 2 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Ordinal Logistic Regression & Score Test

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**proc** **freq** data=lab3;

table cigever4;

**run**;

**proc** **genmod** data=lab3;

class cigever4 (param=ref ref='2');

/\*sets reference group as those who have never had a cigarette\*/

model fritob5 = cigever4/ link=cumlogit dist=mult;

estimate "OR yes vs no" cigever4 **1**/exp;

**run**;

**quit**;

**Proc** **logistic** data=lab3;

class cigever4 (param=ref ref='2');

model fritob5 = cigever4;

**run**;