Example: Poisson Regression

- We analyze the coronary heart disease (CHD) data. The study observed $n = 3,154$ males ages 40-50. The study featured a prospective cohort design, with staggered entry and the observation period concluding on 12/31/70. Men were followed for an average of 8 years, and the number of CHD cases was recorded. Risk factors of interest included smoking, blood pressure and behavior type (A and B).

(a) Read in and print out the analysis file.

See the SAS code.

(b) Fit a main effects model using Poisson regression, but do *not* use an offset. Comment on the validity of such an approach, making reference to the data set.

Y: CHD count

T_i: person year

Model

$$\begin{aligned} log(\mu_i) &= log(\lambda_i) = \beta_0 + \beta_1 I(type_A) + \beta_2 Bp \\ &+ \beta_3 I(cig = 2) + \beta_4 I(cig = 3) \\ &+ \beta_5 I(cig = 4) \end{aligned}$$

This model is not valid since T_is (pearson year) are unequal.

(c) Examine the parameter estimates based on the *no-offset* Poisson model (significance, direction).

Many regression coefficients are significant. It seems like that the directions of bp and smoke effects are not correct.

(d) Under what circumstances would the *no-offset* model be valid?

If $T_i = T$ for all i.

(e) Fit another main effects model, this time using an offset. Test the significance of each term in the model using the Wald test.

Model (offset = $log(T_i)$)

$$\begin{aligned} log(\mu_i) - log(T_i) &= log(\lambda_i) = \beta_0 + \beta_1 I(type_A) + \beta_2 Bp \\ &+ \beta_3 I(cig = 2) + \beta_4 I(cig = 3) \\ &+ \beta_5 I(cig = 4) \end{aligned}$$

BP and smoke directions are corrected.

(f) Interpret $\exp\{\widehat{\beta}_0\}$.

$$\exp\{\widehat{\beta}_0\} = \exp^{-5.47} = 0.0042$$

Estimated CHD rate per person year for a type B non-smoker with BP < 140.

Per person year is a small time unit, so the rate is very small.

(g) Describe the estimated Type A effect, referring to the SAS output.

$$\exp\{\widehat{\beta}_1\} = \exp^{0.76} = 2.14$$

CHD rate ratio between type A and type B, adjusting for other covariates.

(h) Test the null hypothesis of no SMOKE effect, testing all categories simultaneously.

$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$

Test statistic: 24.32

P-value < 0.001

Reject H0

(i) Does the model fit the data well? Carry out GoF.

H0: the model fits data well

Test statistics: D=19.28

Since $D > 18.3 = \chi^2_{10,0.05}$, we can reject H0 at $\alpha = 0.05$

The model does not fit the data well.

The $D/df > 1$, which indicates that there is an over dispersion problem.

(g) Re-fit the model, with PY replaced by $PY/1000$. Compare the parameter estimates to those obtained previously.

New model:

$$\begin{aligned} log(\mu_i) - log(T_i/1000) &= log(\lambda_i^*) = \beta_0^* + \beta_1^* I(type_A) \\ &+ \beta_2^* Bp + \beta_3^* I(cig = 2) \\ &+ \beta_4^* I(cig = 3) + \beta_5^* I(cig = 4) \end{aligned}$$

- Since $\beta_0^* + log(T_i) - log(1000) = \beta_0 + log(T_i)$, $\beta_0^* = \beta_0 + log(1000)$. Therefore,

$$\widehat{\beta}_0^* = \widehat{\beta}_0 + 6.91 = -5.47 + 6.91 = 1.44$$

$$\exp^{1.44} = 4.22$$

- $\beta_i^* = \beta_i$ for all $i = 1, \cdots, 5$