PUBH 7430

On this assignment, as in all assignments, **inclusion of screenshots of R/SAS output is not acceptable**. *Tables and plots prepared from statistical output should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. Obtaining the correct results from an analysis is only a portion of each exercise. It is also very important to write clearly about what your results mean. This includes presenting your results in a way that clearly answers the question and places little burden on the reader.*

You should submit your assignment on Canvas as a PDF file, titled "PubH7430_LastNameFirstName_A2.pdf" with your own name replacing LastNameFirstName. While no code should be included in the body of the assignment, you **must include your code as an appendix** *at the end of the assignment. Keep in mind that while working together on homework assignments is permitted, each student is expected to independently write up homework assignments, including any code, in their own words.*

Background

This dataset is on dental caries (tooth decay) caused by mutans streptococci (MS).

151 persons with gum disease (elevated oral MS levels) were recruited into a study at a mid-western research university and randomly assigned to one of three treatment groups. The treatment under study was a gum with an active drug (Xylitol). There were two control groups: one group (the placebo group) received gum with Sorbitol but no Xylitol and one group received no gum. Participants in the gum groups were asked to chew the gum three times daily for a minimum of 5 minutes each time and carry out their usual oral hygiene (tooth brushing, mouthwash, etc.). Participants in the group without gum were asked to carry out their usual oral hygiene only.

Subjects rinsed their mouths twice daily for 14 days with a 0.12% CHZ gluconate mouthrinse. They were then asked to follow their assigned treatment for three months. Participants had their oral levels of mutans streptococci (MS) measured four times: at baseline after the 14 days of CHZ gluconate mouthrinse (after which they were randomly allocated to one of the three treatment groups), one week later, 1 month later, and 3 months later. The four mut* variables are the levels of oral mutans streptococci at the four time periods in units of log(cfu/ml) (where cfu = "colony forming units", a count of blotches on a standard sized petri dish after standard preparation). Missing MS levels are recorded as 'NA' in the R data file ("gumR.csv" on Canvas) and '.' in the SAS data file ("gumSAS.csv" on Canvas).

The variables dmft and dmfs reflect the general condition of the teeth at baseline: dmft stands for diseased/missing/filled teeth and dmfs stands for diseased/missing/filled surfaces. (There are five surfaces per tooth, the four sides and the chewing surface.)

Questions

1. **Regression and log transformation [18 pts].** Suppose we are interested in exploring the association between condition of the teeth (as reflected by variable *dmfs*) and baseline mutans streptococci levels.

   Note on missing values: *There were several missing values in the dataset, coded as 'NA' in the* `R` *file and '.' in the* `SAS` *file.* **Be aware of how your chosen statistical software package treats missing values!** *For example, many commands in* `R` *(*`mean`*,* `sd`*,* `range`*, etc.) will produce the value 'NA' when working with data containing missing values, but will not warn you that this occurred. The default behavior in regression models (in both* `R` *and* `SAS`*) is to drop any observations containing missing outcomes or covariates; whether or not you are warned that this happened can depend on the function/procedure you are using.*

   (a) [2 pts] Produce histograms showing the distribution of *dmfs* and log(*dmfs*).

   (b) Fit the following four statistical models with *dmfs* as the outcome variable and baseline mutans streptococci level as the predictor. For each model, write down the assumptions you are making about (1) the mean and (2) the variance in terms of $\boldsymbol{\beta}$, and (3) summarize your conclusions about the association between dmfs and baseline MS in a sentence or two suitable for inclusion in a scientific publication. When appropriate, provide an interpretation of the exponentiated coefficient for baseline MS.

      i. [3 pts] A Gaussian GLM with identity link (i.e., a linear regression).

      ii. [3 pts] A Gaussian GLM with log link

      iii. [3 pts] A Gaussian GLM with identity link, using log-transformed values of *dmfs* (Given the zero values, you will need to make a small modification to the data to fit this model. For example, you can replace the zero values for dmfs with 0.01)

      iv. [3 pts] A Poisson GLM with log link (use non-transformed *dmfs* values).

   (c) [2 pts] Comment on the differences you note between the four models fitted in the previous part.

   (d) [2 pts] Produce plots of residuals versus fitted values for each model you fit in 1(b), and comment on any differences/similarities. Do you see evidence of a mean-variance relationship for this outcome? Explain.

2. **Treatment effects** [12 pts] Suppose we are interested in testing the efficacy of the active treatment versus gum placebo at three month post-baseline (outcome *mut3mos*). For questions 2a through 2c, exclude the patients who received no gum.

   (a) [3 pts] Fit a Gaussian GLM with identity link (i.e. a linear regression) to assess the effect of treatment on mutans streptococci levels at three months, and write a sentence summarizing the estimated treatment effect.

   (b) [3 pts] Fit the same model, but also include mutans streptococci level at baseline as a predictor. Write a sentence summarizing the estimated treatment effect from this model. How much, if at all, do your conclusions change?

   (c) [3 pts] Fit a model to assess whether the effect of treatment differs by baseline MS level. Compare the coefficient of the treatment term (i.e., the "main effect" of treatment) in this model to the coefficients estimated in parts (a) and (b), and comment on and explain any differences observed.

1: ii Gaussian Model with a log link.

$$\log(E(Y_i)) = \beta_0 + \beta_1 x_1 \Rightarrow$$

$$E(Y_i) = e^{\beta_0 + \beta_1 x_1}$$

1 unit increase is e^{β_1}

$$\log(E(Y_2)) - \log(E(Y_1)) =$$

$$\log\left(\frac{E(Y_2)}{E(Y_1)}\right) = \beta_0 + \beta_1 X_2 - \beta_0 - \beta_1 X_1 =$$

$$= \beta_1 (X_2 - X_1)$$

$$e^{\log\left(\frac{E(Y_2)}{E(Y_1)}\right)} = e^{\beta_1 (X_2 - X_1)}$$

if $X_2 - X_1 = 1$, then,

$$\frac{E(Y_2)}{E(Y_1)} = e^{\beta_1}$$

Poisson Regression:

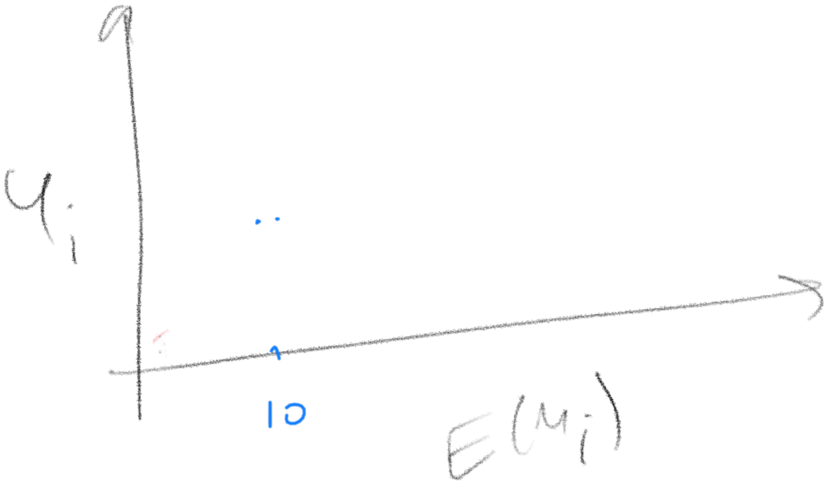
$\log(E(Y_i)) = \beta_0 + \beta_1 x_1$

$E(Y_i) = e^{\beta_0 + \beta_1 x_1}$

for poisson

$E(Y_i) = Var(Y_i)$

$sd = \sqrt{Var Y_i}$



Allow Var

$Var(Y_i | X_i) = E[Y_i](1 - E[Y_i])$

$$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 X_1+\beta_2 x_2},$$

$$p_i = (1-p_i)\, e^{\beta_0 + \beta X_1 + \beta_2 x_2},$$

$$p_1 = e^{X\beta} - p_i e^{x\beta},$$

$$e^{x\beta} = p_i + p_i e^{x\beta}$$

$$e^{x\beta} = p_i(1+e^{x\beta}),$$

$$p_i = \frac{e^{x\beta}}{1+e^{x\beta}}.$$

$$1-p_i = 1 - \frac{e^{x\beta}}{1+e^{x\beta}} = $$

$$= \frac{1+e^{x\beta} - e^{x\beta}}{1+e^{x\beta}} = \frac{1}{\cdots}$$

$$\log\left(\frac{P_2}{1-P_2}\right) - \log\left(\frac{P_1}{1-P_1}\right) =$$

$$= \beta_2 - \beta_1.$$

$$\log\left[\frac{\frac{P_2}{1-P_2}}{\frac{P_1}{1-P_1}}\right] = \beta_2 - \beta_1, \quad \frac{\frac{P_2}{1-P_2}}{\frac{P_1}{1-P_2}} = e^{\beta_2-\beta_1}$$

$$\log(p_2) = \beta_0 + \beta_1 X_2 ,$$

$$\log(p_1) = \beta_0 + \beta_1 X_1$$

$$\log(p_2) - \log(p_1) = \beta_1 (X_2 - X_1),$$

$$\log\left(\frac{p_2}{p_1}\right) = \downarrow ,$$

$$\frac{p_2}{p_1} = e^{\beta_1 (X_2 - X_1)} = e^{\beta_1}$$

$$\boxed{se(\beta_2 - \beta_1) = Var(\beta_2) + Var(\beta_1) - 2\,Cov(\beta_2, \beta_1)}$$

- Assume that X_2, X_1 are fixed.

- Odds Ratio:

$$\frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}} = \frac{p_2}{1-p_2} \times \frac{1-p_1}{p_1}$$

(d) [3 pts] Fit a model similar to the model described in 2(b) but this time compare the group that was given the placebo gum to the group that was not given gum. Does there appear to be a "placebo effect" whereby subjects receiving the placebo gum have better outcomes than those who did not receive gum?

3. **Treatment effects, part 2.** [10 pts] Suppose that your colleagues in the Dental School tell you that having a mutans streptococci level greater than 4.5 log(cfu/ml) is known to be associated with particularly poor dental outcomes.

(a) [4 pts] Define a binary outcome variable indicating whether or not a patient's MS level exceeds 4.5 log(cfu/ml) at one month post-baseline. Fit a model to compute the odds ratio of this outcome associated with treatment (Xylitol vs. gum placebo). Write down the assumptions you are making about (1) the mean and (2) the variance in terms of $\boldsymbol{\beta}$. Summarize your conclusions in a sentence or two suitable for inclusion in a scientific publication.

(b) [4 pts] Using the binary outcome variable defined in 4(a), now fit a model to compute the relative risk of this outcome associated with treatment (Xylitol vs. gum placebo). Write down the assumptions you are making about (1) the mean and (2) the variance in terms of $\boldsymbol{\beta}$. Summarize your conclusions in a sentence or two suitable for inclusion in a scientific publication.

(c) [2 pts] Would you expect the results from 4(a) and 4(b) to be similar? If not, when would you expect similar results?