BIOS 6612 Homework 3: Case-control data and other link functions for binary response data

Solutions

1. (**30 points**) Bacterial growth in donated blood can lead to severe infections in transfusion patients. In the absence of limiting factors, bacteria growth follows a power law, with a strain-specific doubling time of θ in some arbitrary time unit.

   Suppose we have $n = 20$ bags of donated blood. Each contains 100 mL of blood and is infused with a common strain of bacteria at time 0. At $t = 8, 12, 16$ hours subsequent to time 0, a 1-mL sample is drawn from each bag and tested for the presence or absence of bacteria. The table below shows the number of bags testing positive at each time point.

Time (hours)	Num. bags positive
8	5
12	10
16	19

   Assume that the actual number of bacteria present in a sample of (known) fraction k at time t follows a Poisson distribution with mean $k \cdot 2^{t/\theta}$. Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})'$ be the indicators of presence ($Y_{ij} = 1$) or absence ($Y_{ij} = 0$) of bacteria at each of the three observation times for the ith blood bag.

   (a) Give the likelihood for the data as a function of θ. (**8 points**)
   From the statement of the problem we have $Z_{ij} \sim \text{Poi}(k \cdot 2^{t_j/\theta}), t_j \in \{8, 12, 16\}$. We need to write the likelihood as a function of the observed data, which is $Y_{ij} = \mathbb{1}(Z_{ij} > 0)$ (2 points). This is a Bernoulli random variable, so all we need to do is find its mean to write out the likelihood. From the Poisson mass function, we know that

$$P(Z_{ij} > 0) = 1 - \exp(-k \cdot 2^{t_j/\theta}) \text{ (2points)}.$$

   Therefore, the likelihood of the data in subject-level form is

$$L(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{3} \left\{1 - \exp(-k \cdot 2^{t_j/\theta})\right\}^{Y_{ij}} \left\{\exp(-k \cdot 2^{t_j/\theta})\right\}^{1-Y_{ij}}.$$

Grouped form is also acceptable, and given by (with data from the problem substituted in)

$$\begin{aligned} L(\theta) \propto & \left[\left\{1-e^{-k\cdot 2^{8/\theta}}\right\}^{5}\left\{e^{-k\cdot 2^{8/\theta}}\right\}^{20-5}\right] \\ & \times\left[\left\{1-e^{-k\cdot 2^{12/\theta}}\right\}^{10}\left\{e^{-k\cdot 2^{12/\theta}}\right\}^{20-10}\right] \\ & \times\left[\left\{1-e^{-k\cdot 2^{16/\theta}}\right\}^{19}\left\{e^{-k\cdot 2^{16/\theta}}\right\}^{20-19}\right] \end{aligned}$$

For grouped form here, the likelihood is proportional ($\propto$) because the binomial coefficient terms $\binom{20}{5}$, $\binom{20}{10}$, $\binom{20}{19}$ have been dropped, and would be equal if these had been included. (4 points for the correct likelihood in grouped or ungrouped form).

(b) Show that the MLE of θ can be obtained from a GLM with a binary response, covariate of t_j, complementary log-log link function, offset equal to $\log k$, and no intercept. (**8 points**)
We see from the form of the likelihood that we have a binary response model. We just need to rewrite the mean in GLM form to see that each of the components matches up with what we are told should be there in the question: The form of a GLM is

$$g(\mathbb{E}(Y|t)) = \eta \text{ (1 point)}$$

$$\begin{aligned} \mathbb{E}(Y|t) &= 1-\exp(-k\cdot 2^{t/\theta}) \\ &= 1-\exp\left\{-\exp\left(\log k + t/\theta\cdot\log 2\right)\right\}. \end{aligned}$$

Let $p = \mathbb{E}(Y|t)$ and let $\eta = \log k + t/\theta\cdot\log 2$ (1 point). Then we have

$$\begin{aligned} p &= 1-\exp\{-\exp(\eta)\} \\ 1-p &= \exp\{-\exp(\eta)\} \\ \log(1-p) &= -\exp(\eta) \\ \log\{-\log(1-p)\} &= \eta \text{ (2 points).} \end{aligned}$$

On the left-hand side is the form of the complementary log-log link. For the linear predictor, we have

$$\eta = \log k + t/\theta\cdot\log 2 = \log k + \frac{\log 2}{\theta}t.$$

There is no intercept here; instead we have a term $\log k$ with a known coefficient (of 1) which does not have to be estimated, corresponding to the offset. The coefficient $\beta_1 = (\log 2)/\theta$ (4 points) multiplies our covariate t, completing the specification of our GLM.

(c) Fit this model using standard software for generalized linear models; provide your code and model output. Note that $k = 1/100$ since each sample is 1 mL drawn from the original volume of 100 mL in each bag. (**8 points**)

```
> library(tidyverse)
>
> growth = tibble(
+   time = c(8, 12, 16),
+   positive_bags = c(5, 10, 19),
+   n = rep(20, 3),
+   k = 1/100
+ )
>
> growth_mod = glm(cbind(positive_bags, n - positive_bags) ~ 0 + time +
 offset(log(k)),
+                  family= binomial(link = "cloglog"),
+                  data = growth)
>
> summary(growth_mod)

Call:
glm(formula = cbind(positive_bags, n - positive_bags) ~ 0 + time +
    offset(log(k)), family = binomial(link = "cloglog"), data = growth)

Deviance Residuals:
      1        2        3
 1.0021  -0.2253  -0.1380

Coefficients:
     Estimate Std. Error z value Pr(>|z|)
time  0.35921    0.01593   22.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 255.8514  on 3  degrees of freedom
Residual deviance:   1.0739  on 2  degrees of freedom
AIC: 11.691

Number of Fisher Scoring iterations: 4
```

(2 points for correct parameter estimate, 2 points for correct specification of offset, 2 points for correct specification of cloglog link function, 2 points for correct specification of outcome in model formula)

(d) Using your model output, give the MLE of θ and an associated 95% confidence interval. (**6 points**)
From our answer above placing this model in GLM form, we know that $\beta_1 =$

$(\log 2)/\theta$. Let $g(\beta_1) = (\log 2)/\beta_1$. Then the MLE of θ is $g(\hat{\beta}_1) = (\log 2)/\hat{\beta}_1 = 1.9296$ (3 points). We can get the confidence interval using the delta method: $g'(\beta_1) = -(\log 2)/\beta_1^2$, so $\text{Var}(\hat{\theta}) \approx \text{Var}(\hat{\beta}_1)\{-(\log 2)/\beta_1^2\}^2$ so the SE of $\hat{\theta}$ is $0.01593 \times 0.6931/0.35921^2 = 0.08557$. The associated 95% confidence interval for $\hat{\theta}$ is therefore $\hat{\theta} \pm 1.96 \times \text{SE}(\hat{\theta}) = (1.7619, 2.0973)$ (3 points). Alternatively, we can get the confidence interval by transforming the endpoints on the linear predictor scale (3 points, showing work).

(e) Compute a measure of model deviance that can be approximated by a χ^2 distribution and assess the fit of the model. (**5 points extra credit**)
We can use the residual deviance from the above model fit since this model was fit to the data in grouped form. The test statistic is 1.0739 which we can compare to a χ^2_2 reference distribution: this gives a p-value of approximately 0.58, so we fail to reject the null and conclude that this model is an adequate fit to the data.

2. (**40 points**) We can use logistic regression to estimate the odds ratio as a measure of association between a binary exposure and a binary outcome regardless of whether cohort or case-control sampling is used. For this question, you will need to conduct a simulation study to verify this result. The parameters needed for this simulation are given below:

- Population size N =100,000
- Prevalence of disease
  - 5% in unexposed
  - 10% in exposed
- Prevalence of exposure is 30%

(a) Give the parameters needed to simulate the outcome conditional on exposure, assuming a logistic regression model in the population. (**6 points**)
To answer this, we only need the prevalence of disease (i.e., $P(Y = 1)$ in exposed and unexposed (3 points each, -1 if extra information provided). If we define the exposure variable $X = 0$ for unexposed and $X = 1$ for exposed, then $P(Y = 1|X = 0) = 0.05$ and $P(Y = 1|X = 1) = 0.10$. Our logistic regression model is

$$\log \frac{P(Y = 1|X)}{P(Y = 0|X)} = \beta_0 + \beta_1 X,$$

so β_0 is the logit of the probability of disease among the unexposed, $\beta_0 = \log(0.05/0.95) \approx -2.94$. We can find β_1 as the log odds ratio for disease among the exposed compared to the unexposed, or

$$\beta_1 = \log \frac{0.1/0.9}{0.05/0.95} \approx 0.75.$$

(b) Generate data on exposure and outcome for the entire population. *Hint:* In R, you can use the function `rbinom(N,1,prob)` to generate `N` independent Bernoulli variables with mean `prob`. (**9 points**)
Below we generate a population with some randomness, based on the given population-level parameter values. We could also generate this population deterministically.

```
## population level parameters
p_d_exposed = 0.1  # P(Y = 1 | X = 1)
p_d_unexposed = 0.05 # P(Y = 1 | X = 0)
p_exposure = 0.3
N = 100000

## always set a seed!
set.seed(20202021)

## generate the X values
x = rbinom(n = N, size = 1, prob = p_exposure)
beta0 = log(p_d_unexposed/ (1- p_d_unexposed))
beta1 = log(p_d_exposed / (1 - p_d_exposed) / (p_d_unexposed / (1- p_d_unexpose

# get probabilities from population parameters based on logistic model
p_yGivenx = plogis(beta0 + beta1 * x) # inverse logit function

## generate the Y|X values
y = rbinom(n = N, size = 1, prob = p_yGivenx)

# put it into a data frame / tibble
population_data = tibble(
  subject_id = 1:N,
  y = y,
  x = x
)
```

(c) Now that you have a population, conduct 1000 simulations where you randomly select $n = 50$ individuals with the exposure and another $n = 50$ without the exposure; this represents a cohort study. For each simulation, fit the appropriate logistic regression model and record the parameter estimates. Give both the mean and median of each the estimates of each parameter across all simulations. Why might the estimates not agree with one another across simulated samples? How do they compare to the true estimate values, and why might they differ? (**10 points**)

```
# define function that gets estimated beta values for a specific seed and subsa
resample_logistreg = function(seed, sample_size){
```

```
  set.seed(seed)

  sample_data = population_data %>%
    group_by(x) %>%
    slice_sample(n = sample_size) %>%
    ungroup()

  logistic_mod = glm(y ~ x,
                     family = binomial,
                     data = sample_data)

  # put estimated coefficients into a dataframe and return this data frame
  estimates = tibble(
    seed = seed,
    beta0_hat = as.numeric(coef(logistic_mod)[1]),
    beta1_hat = as.numeric(coef(logistic_mod)[2])
  )

  return(estimates)
}

estimated_coefs_cohort = map_dfr(100001:101000,
.f = resample_logistreg, sample_size = 50) # (2 points for correct sample size)

estimated_coefs_cohort = estimated_coefs_cohort %>%
  mutate(beta0_true = beta0,
         beta1_true = beta1)

estimated_coefs_cohort %>%
+   summarize(mean_est_beta0 = mean(beta0_hat),
+             median_est_beta0 = median(beta0_hat),
+             true_beta0 = first(beta0),
+             mean_est_beta1 = mean(beta1_hat),
+             median_est_beta1 = median(beta1_hat),
+             true_beta1 = first(beta1))
# A tibble: 1 x 6
  mean_est_beta0 median_est_beta0 true_beta0 mean_est_beta1 median_est_beta1 tr
           <dbl>            <dbl>      <dbl>          <dbl>            <dbl>
1          -4.23            -3.18      -2.94           1.85            0.759
```

Your answers from one simulation to another will differ slightly due to a different random seed. One reason these estimates could be different from the true population level values is that our sample size is not large enough for asymptotic normality to kick in. This could be the case due to the low rate of disease in both exposed and unexposed individuals. (4 points for explanation)

(d) Repeat the previous question for a case-control design: that is, randomly sample $n = 50$ individuals with the disease and $n = 50$ without the disease, then fit the appropriate logistic regression model. As before, record the estimated values of the parameters for each of your 1000 simulations. Give the mean and median of the estimates across all simulations. (**8 points**)

```
resample_logistreg_cc = function(seed, sample_size){
+   set.seed(seed)
+
+   sample_data = population_data %>%
+     group_by(y) %>%
+     slice_sample(n = sample_size) %>%
+     ungroup()
+
+   logistic_mod = glm(y ~ x,
+                      family = binomial,
+                      data = sample_data)
+
+   # put estimated coefficients into a dataframe and return this data frame
+   estimates = tibble(
+     seed = seed,
+     beta0_hat = as.numeric(coef(logistic_mod)[1]),
+     beta1_hat = as.numeric(coef(logistic_mod)[2])
+   )
+
+   return(estimates)
+ }
>
> estimated_coefs_cc = map_dfr(100001:101000,
.f = resample_logistreg_cc, sample_size = 50)
>
> estimated_coefs_cc %>%
+   summarize(mean_est_beta0 = mean(beta0_hat),
+             median_est_beta0 = median(beta0_hat),
+             true_beta0 = first(beta0),
+             mean_est_beta1 = mean(beta1_hat),
+             median_est_beta1 = median(beta1_hat),
+             true_beta1 = first(beta1))
# A tibble: 1 x 6
  mean_est_beta0 median_est_beta0 true_beta0 mean_est_beta1 median_est_beta1 tr
           <dbl>            <dbl>      <dbl>          <dbl>            <dbl>
1         -0.271           -0.268      -2.94          0.746            0.754
```

(e) Both the cohort and case-control designs are estimating the same odds ratio between exposure and disease, but their efficiency might be different. Compare the

variability of the estimated log odds ratios between the two designs and explain any differences you find. *Hint:* the `mad()` command in R computes a robust estimate of spread and might be useful if you find large differences between mean and median estimated parameter values. (**7 points**)

```
> sd(estimated_coefs_cohort$beta1_hat)
[1] 4.773774
> sd(estimated_coefs_cc$beta1_hat)
[1] 0.4312162
> # numerical problems result from small sample size
> # combined with low prevalence of the outcome,
> # so we can instead compare MAD (median absolute deviation)
> mad(estimated_coefs_cohort$beta1_hat)
[1] 0.8949844
> mad(estimated_coefs_cc$beta1_hat). # (2 points for producing estimates)
[1] 0.401388
```

We see greater variability in the estimated log odds ratio under the cohort design than under the case-control design (2 points). Think about the 2×2 tables that result from each of these designs: with the cohort design, the cell counts are much smaller in the $Y = 1$ row of the table compared with those in the case-control design. Since the approximate variance of the log odds ratio is equal to $1/a + 1/b + 1/c + 1/d$ and this is minimized for a fixed value of $n = a + b + c + d$ with $a = b = c = d = n/4$, the larger deviation from this configuration associated with the cohort design implies that its estimates of the odds ratio will be less precise (1 point).