Medical Statistics Assignment

James Aley

November 2015

1 Analysis of a Journal Article

This article discusses statistical elements of the analyses in a medical study focusing on the impact coffee consumption may have on risk of individuals later suffering from type 2 diabetes mellitus [1].

The study is a prospective analysis, using individuals selected randomly from public registers in two Dutch towns. Participants were sent a questionnaire to record various aspects of their lifestyle considered relevant to the study. Importantly, this of course covered how much coffee the participants consumed, measured in cups/day, but also various other factors that were thought to be possibly confounding in the study.

We note at this point that this is necessary, as though the study is prospective, it is *not* a clinical trial. We do have an element of randomization to help control for confounds, but the explanatory variable of interest, the amount of coffee consumed by each individual, has not been actively manipulated as a part of the study. The study therefore can only hope to demonstrate a strong association between coffee consumption and risk of type 2 diabetes. We will see later, however, that some considerable effort is made to monitor the impact of potentially confounding factors.

The study excludes a number of individuals upfront as part of the design. This includes people that are under 30 years old, pregnant, or had missing information in the survey. After these exclusions, the total number of participants was 17,111. These individuals were sent follow-up questionnaires over subsequent years to check whether they had been diagnosed with type 2 diabetes. In total, 306 such cases were recorded in the study.

A 95% confidence interval for incidence of type 2 diabetes within this population would therefore be (0.016, 0.020), with estimated value 0.017. Other studies placed overall population incidence for the Netherlands at 2.2% in 2003 [2]. It seems likely that the age restrictions on this study (30-60 years) may explain this slightly lower observation in the study cohort. However, we would also expect there to be some selection bias, in that participants diagnosed with diabetes may be more enthusiastic to respond to follow-ups than others.

The article covers some measures used to monitor and control for potential confounds in the analysis. This seems like a good idea, as coffee consumption is

likely to correlate quite strongly with various other lifestyle choices that could have an impact. Many such potential confounding factors are identified in the study, but we should note that this highlights one of the main drawbacks in this kind of study. In order to test for confounds, we have to propose and enumerate them. It is entirely plausible that others exist, that we're unable to foresee at the beginning of the study, when designing the questionnaire for participants to complete. We therefore would have no data about such factors and be unable to control for them when interpreting the results.

The article states that for each of the characteristics they measured in participants, tests for correlation with coffee consumption were "obtained by modeling the median value of each category of coffee consumption as a continuous variable". It's not entirely clear what is meant by this, as no further details are provided in other parts of the article, however the need to map consumption categories to a continuous variable perhaps suggests some kind of regression analysis is being used.

These tests find that coffee consumption correlates somewhat with all characteristics except for cardiovascular disease and consumption of dark bread. One of the more surprising elements of this study is that those characteristics showing positive correlation with coffee consumption are generally considered to have a negative impact on health. If we were to attribute the conclusions in this study to the confounding factors considered here, it would perhaps be even more surprising than attributing it to coffee consumption anyway.

The study makes use of the *Cox proportional hazards model* [3] to estimate relative risk for participants as a function of their coffee consumption, time and other factors suspected to be confounding. This is a time-series analysis technique, which seems like a sensible choice, given that we cannot ignore how time plays an important role in the diagnosis of diseases such as type 2 diabetes. We do not have an immediately measurable response variable; all we know is that a participant has or has not *yet* been diagnosed. Censoring therefore plays an important role in interpreting the results.

The Cox proportional hazards model relies on the *proportional hazards assumption*, given below:

$$h_z(t) = g(z)h_0(t) \tag{1}$$

Where z is a vector of explanatory variables, $h_0(t)$ is the baseline or nominal hazard function and $g(z)$ is a function of *only* the explanatory variables (not time), which we multiply by the baseline hazard rate to obtain the hazard function taking z into account. Thus we assume that the impact of z does not vary with time beyond the base hazard rate.

It's very hard to check whether this assumption is reasonable, as we don't know what (if anything) in coffee could be helping to prevent type 2 diabetes. We certainly have no information to tell us whether such an effect will vary with time, whether it be how long a participant has been drinking coffee at the specified quantity, their age or length of exposure to potential causes.

Proceeding with the Cox regression model, however, we can reasonably as-

sume that the relative risk estimates cited in the article have been produced using a model of the form:

$$\frac{h_z(t|X_1,\ldots,X_k)}{h_0(t)} = \exp(\beta_1 X_1 + \ldots + \beta_k X_k) \tag{2}$$

The ratio of h_z to h_0 would be the relative risk reported, and the $X_1 \ldots X_k$ are explanatory variables for the regression model. In this case, the explanatory variable of primary interest would be coffee consumption levels. Again, the article mentions that consumption categories were mapped to continuous variables with the same technique as before.

The conclusions reached about the relative risk for high consumers of coffee being approximately 0.50 are based on inferences around the corresponding β value in the regression model. If we let β_c be our coefficient corresponding to the continuous random variable created for coffee consumption in cups/day, then the p-values provided are presumably testing the null hypothesis $H_0 : \beta_c = 0$ vs the alternative $H_1 : \beta_c \neq 0$. The very low p-value indicates that level of coffee consumption significantly improves the model for predicting diagnoses over time.

It is useful to consider at this point what the relative risk, or hazard ratio in this model actually represents, and how the response variable was initially measured in the study. The hazard function itself is the probability of an event occurring after time t, *in the next instant*. In this study, that probability represents diagnosis of type 2 diabetes. However, such a diagnosis is not really an immediate effect, it requires that a participant notices symptoms in themselves, visits a medical professional and then later completes a questionnaire providing the date this happened. Many patients first go through phases of being "pre-diabetic" before their conditions worsen and a diagnosis of full type 2 diabetes is issued.

The requirement for action from the patient in order to give us a value for t could pose problems. It was already demonstrated in the analysis of potential confounding factors that coffee consumption correlates quite strongly with various other (mostly unhealthy) life-style choices. It seems entirely possible that participants drinking more coffee may also visits doctors less frequently, on average. They may also be less likely to perceive problems in the symptoms, as they are initially quite subtle (increased thirst and hunger, for example). The study makes no mention of this, or of any requirement for participants in the study to regular have checkups, which may have helped control for issues like this.

Though we can see that there are various parts of the study susceptible to systematic problems, the overall conclusions are certainly interesting. We can see that there are a few points where the statistical analysis may have been somewhat sub-optimal, but for such a strong measured effect, it seems unlikely that it would impact conclusions substantially. The biggest risk seems to be that perhaps the very design of the study, relying on action from participants to provide a value for the time until diagnosis, may actually correlate with the

same personality traits that the study demonstrated coffee consumption does. The methods used to control for confounds in this study do not control for that particular source of potential bias. It may be interesting to repeat the analysis with logistic regression, modeling the proportion of cases in a fixed amount of time, rather than as a function of time. If the hypothesis about diagnosis time correlating with consumption is correct, we would expect to see a smaller effect this way.

2 Newton-Raphson for Binomial Max-Likelihood

The code in the following listing provides an implementation for the Newton-Raphson algorithm, to fit parameters that maximize the likelihood function for the Binomial distribution. The code can also be access online, see [4].

```
#### Newton-Raphson implementation for logistic
    regression
#### Medical Statistics Assignment, MSc Applied
    Statistics
#### James Aley

# The 'infert' dataset contains data about infertility
    after induced and
# spontaneous abortions. The data are given with a binary
     0/1 response for
# cases vs controls.

# To start with, let's use R's implementation in the GLM
    function with this
# dataset, so that we have a golden answer to compare our
     algorithm to.

reference.model <- glm(case ~ factor(induced) + factor(
    spontaneous),
                       data = infert,
                       family = "binomial")

reference.model$coefficients

### Our implementation of the N-R algorithm should
    produce coefficients
### quite similar to these, although we note that R is
    actually using
### a different algorithm (Fisher Scoring)

#  (Intercept)      factor(induced)1      factor(induced)2
    factor(spontaneous)1
```

```
# -1.7442332                0.4608354                0.8249893
                  1.2892855
# factor(spontaneous)2
# 2.3537835


### To implement the N-R algorithm, we will need some
    extra functions:
### The sigmoid function for the logit link, plus the
    first and second
### derivatives of the binomial log-likelihood function.

#' Sigmoid function: e^x / 1+e^x
sigmoid <- function(x) {
  (exp(x) / (1 + exp(x)))
}

#' First derivative of the binomial log-likelihood
    function.
#' Data should be a matrix where each row has the form:
#'  X, N, <explanatory variables>
#'
#' @param beta the point (vector) to calculate the
    derivative at
#' @return a vector for the first derivative calculated
    at beta
binom.log.likelihood.prime <- function(data, beta) {
  p = length(beta)
  s = matrix(rep(0, p), ncol = 1, nrow = p)
  for(i in 1:nrow(data)) {
    r = data[i, ]
    a <- r[1]
    n <- r[2]
    x <- tail(r, n=-2)

    s = s + (x %*% (a - (n * sigmoid(t(x) %*% beta))))
  }

  s
}

#' Second derivative of the binomial log-likelihood
    function.
#' Data should be a matrix where each row has the form:
#'  X, N, <explanatory variables>
#'
```

```
#' @param beta the point (vector) to calculate the second
     derivative at
#' @return a matrix containing the second derivative at
    point beta
binom.log.likelihood.prime2 <- function(data, beta) {
  p = length(beta)
  s = matrix(rep(0, p*p), ncol = p, nrow = p)
  for(i in 1:nrow(data)) {
    r = data[i, ]
    a <- r[1]
    n <- r[2]
    x <- tail(r, n=-2)

    s = s + (n * sigmoid(t(x) %*% beta)[,1]
             * (1 - sigmoid(t(x) %*% beta))[,1]
             * (x %*% t(x)))
  }

  -1 * s
}

#' Implementation of the Newton-Raphson algorithm.
#'
#' @param x0 A starting value for the algorithm
#' @param data A matrix where each row is: X, N, <
    explanatory variables>
#' @param ll.prime First derivative of log-likelihood
    function
#' @param ll.prim2 Second derivative of log-likelihood
    function
#' @param max.iter Maximum number of iterations to try
    before returning
#' @return The optimised vector of coefficients
newton.raphson <- function(x0, data, ll.prime, ll.prime2,
    max.iter=100) {
  b = x0
  iter = 0
  for(i in 1:max.iter) {
    first.deriv = ll.prime(data, b)
    vnorm = norm(first.deriv, "f")
    if(vnorm < 0.00001) {
      print(sprintf("Converged after %d iterations", i))
      return(b)
    } else {
      second.deriv = ll.prime2(data, b)
      b = b - solve(second.deriv) %*% first.deriv
```

```
    }
  }

  warning("Algorithm did not converge within max
      iterations.")
  return(b)
}

### Now let's test the algorithm:

# Set up some test data, we'll use the design matrix from
# reference.model, so that we know what coefficients the
    algorithm
# *should* produce if it's working properly:

# > d <- cbind(infert$case, rep(1, nrow(infert)),
#               model.matrix(reference.model))

# > newton.raphson(c(1, 0, 0, 0, 0), d,
#                  binom.log.likelihood.prime,
#                  binom.log.likelihood.prime)
#
# [1] "Converged after 5 iterations"
#                                [,1]
# (Intercept)           -1.7442332
# factor(induced)1       0.4608354
# factor(induced)2       0.8249893
# factor(spontaneous)1   1.2892855
# factor(spontaneous)2   2.3537835

### These are very similar to the values given in the
    model we fitted
### with the GLM function above:

# > newton.raphson(c(1, 0, 0, 0, 0), d,
#                  binom.log.likelihood.prime,
#                  binom.log.likelihood.prime)
#   - reference.model$coefficients
#
# [1] "Converged after 5 iterations"
#                                  [,1]
# (Intercept)             2.141562e-08
# factor(induced)1       -4.881931e-09
# factor(induced)2       -1.382456e-08
# factor(spontaneous)1   -1.876198e-08
# factor(spontaneous)2   -2.065155e-08
```

References

[1] Rob M van Dam, Edith J M Feskens. *Coffee Consumption and risk of type 2 diabetes mellitus.* Lancet 2002; 360: 147778

[2] Department of Family Practice, University of Groningen. *Prevalence, incidence and mortality of type 2 diabetes mellitus revisited: a prospective population-based study in The Netherlands (ZODIAC-1).* Eur J Epidemiol. 2003;18(8):793-800.

[3] Cox, D. R., and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London, New York.

[4] James Aley on GitHub. *Newton-Raphson assignment source code.* `https://github.com/jaley/stats-assignments/blob/master/medical_stats/assignment_newton_raphson.R`