Populations, Cox models, and "type III" tests

Terry M Therneau
Mayo Clinic

July 1, 2015

Contents

1 Introduction

This note started with an interchange on the R-help. A user asked "how do I do a type III test using the Cox model", and I replied that this was not a well defined question. If he/she could define exactly what it was that they were after, then I would look into it. To which the response was that "SAS does it". A grant deadline was looming so the discussion did not get any further at that point, but it eventually led to a much longer investigation on my part, which is summarized in this note. There are three central ideas as it turns out: populations, computation, and the mapping linear models ideas onto the Cox model.

The first idea, and perhaps the central one, is using the model fit from a current data set to predict for a new population. This plays an important role in predicted survival curves, see for instance the vignette on that topic or chapter 10 of our book [13]; recognizing that "type 3" tests are simply another variant on that theme was a pivotal step in my understanding. This immediately leads to the important subtopic of "prediction for *which* population". The SAS type 3 computations corresponds to a very particular and inflexible choice.

The second theme is computational: given some summary measure and a population for which you wish to predict it, the result will be some sort of weighted average. There are two primary ways to set up this computation. In a linear model one of them can be reduced to a particular contrast $C\hat{\beta}$ in the fitted coefficients $\hat{\beta}$, which is an appealing choice since follow-up computations such as the variance of the estimate become particularly simple. A common, simple, but unreliable algorithm for creating C has been a major source of confusion (hereafter referred to as the NSTT: not safe type three).

The last theme is how the linear models formulae map to the Cox model case. In particular, there is a strong temptation to use $C\hat{\beta}$ with C taken from linear models machinery and $\hat{\beta}$ from a fitted Cox model. The problem is that this implicitly requires a replacement of $E[\exp(X)]$ with $\exp(E[X])$. For a Cox model $C\beta$ is certainly a valid statistic for any C, we just have no clear idea of what it is testing.

For the impatient readers among you I'll list the main conclusions of this report at the start.

- SAS type 3 predicts for a population with a uniform distribution across all categorical predictors. Scholarly papers discussing fundamental issues with using such an approach as a default analysis method have appeared almost biannually in the statistics literature, with little apparent effect on the usage of the method. SAS documentation of type 3 is almost entirely focused on the algorithm they use for computing C and ignores the population issue.

- Population predictions very often make sense, including the question the type 3 approach is attempting to address. There are valid ways to compute these estimates for a Cox model, they are closely related the inverse probability weight (IPW) methods used in propensity scores and marginal structural models.

- The algorithm used to compute C by the SAS glm procedure is sophisticated and reliable. The SAS phreg procedure uses the linear models approach of $C\hat{\beta}$ to compute a "type 3" contrast, with C computed via the NSTT. The combination is a statistical disaster. (This is true for SAS version 9.4; I will update this note if things change.)

2 Linear approximations and the Cox model

One foundation of my concern has to do with the relationship between linear models and coxph. The solution to the Cox model equations can be represented as an iteratively reweighted least-squares problem, with an updated weight matrix and adjusted dependent variable at each iteration, rather like a GLM model. This fact has been rediscovered multiple times, and leads to the notion that since the last iteration of the fit *looks* just like a set of least-squares equations, then various least squares ideas could be carried over to the proportional hazards model by simply writing them out using these final terms.

In practice, sometimes this works and sometimes it doesn't. The Wald statistic is one example of the former type, which is completely reliable as long as the coefficients β are not too large[1]. A counter example is found in two ideas used to examine model adequacy: adjusted variable plots and constructed variable plots, each of which was carried over to the Cox model case by reprising the linear-model equations. After a fair bit of exploring I found neither is worth doing [13]. Copying over a linear models formula simply did not work in this case.

3 Data set

We will motivate our discussion with the simple case of a two-way analysis. The `flchain` data frame contains the results of a small number of laboratory tests done on a large fraction of the 1995 population of Olmsted County, Minnesota aged 50 or older [4, 2]. The R data set contains a 50% random sample of this larger study and is included as a part of the survival package. The primary purpose of the study was to measure the amount of plasma immunoglobulins and its components. Intact immunoglobulins are composed of a heavy chain and light chain portion. In normal subjects there is overproduction of the light chain component by the immune cells leading to a small amount of *free light chain* in the circulation. Excessive amounts of free light chain (FLC) are thought to be a marker of disregulation in the immune system. Free light chains have two major forms denoted as kappa and lambda, we will use the sum of the two.

An important medical question is whether high levels of FLC have an impact on survival, which will be explored using a Cox model. To explore linear models we will compare FLC values between males and females. A confounding factor is that free light chain values rise with age, in part because it is eliminated by the kidneys and renal function declines with age. The age distribution of males and females differs, so we will need to adjust our simple comparison between the sexes for age effects. The impact of age on mortality is of course even greater and so correction for the age imbalance is is critical when exploring the impact of FLC on survival.

Figure 1 shows the trend in free light chain values as a function of age. For illustration of linear models using factors, we have also created a categorical age value using deciles of age. The table of counts shows that the sex distribution becomes increasingly unbalanced at the older ages, from about 1/2 females in the youngest group to a 4:1 ratio in the oldest.

```
> library(survival)
> library(splines)
> age2 <- cut(flchain$age, c(49, 59, 69, 79, 89, 120),
```

[1] In practice failure only occurs in the rare case that one of the coefficients is tending to infinity. However, in that case the failure is complete: the likelihood ratio and score tests behave perfectly well but the Wald test is worthless.

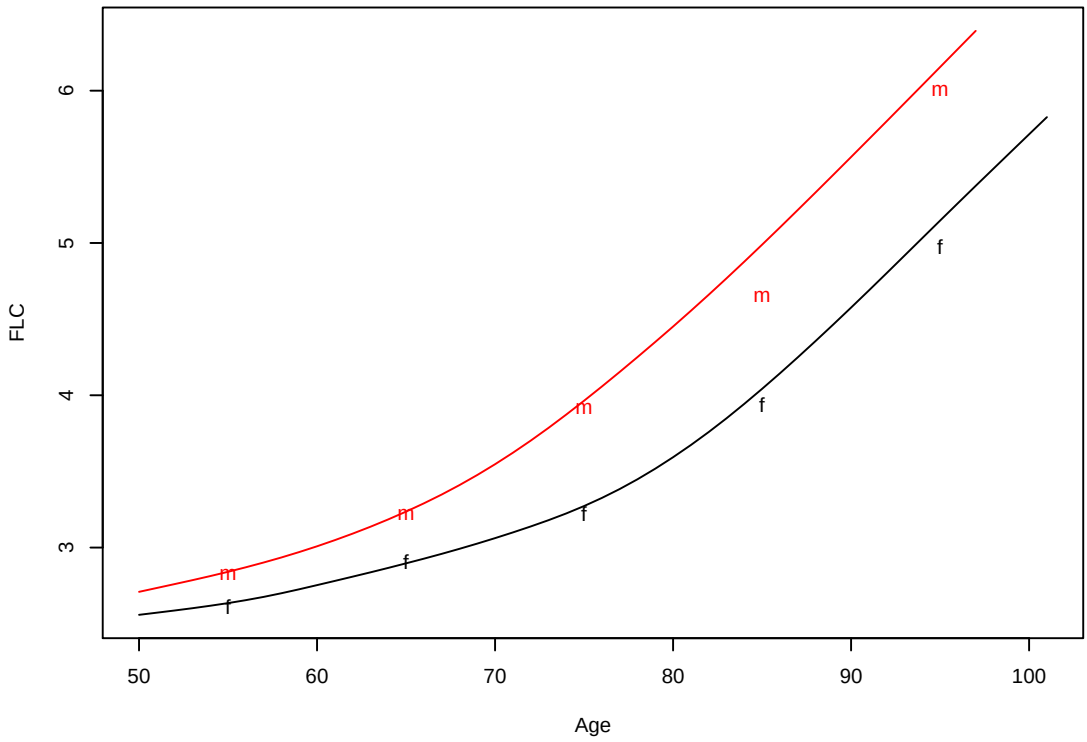

Figure 1: Average free light chain for males and females. The figure shows both a smooth and the means within deciles of age.

```
              labels=c("50-59", "60-69", "70-79", "80-89", "90+"))
> counts <- with(flchain, table(sex, age2))
> counts
   age2
sex 50-59 60-69 70-79 80-89  90+
  F  1647  1214   949   459   81
  M  1510  1115   674   202   23
> #
> flchain$flc <- flchain$kappa + flchain$lambda
> male <- (flchain$sex=='M')
> mlow <- with(flchain[male,],  smooth.spline(age, flc))
> flow <- with(flchain[!male,], smooth.spline(age, flc))
> plot(flow, type='l', ylim=range(flow$y, mlow$y),
       xlab="Age", ylab="FLC")
> lines(mlow, col=2)
> cellmean <- with(flchain, tapply(flc, list(sex, age2), mean, na.rm=T))
> matpoints(c(55,65,75, 85, 95), t(cellmean), pch='fm', col=1:2)
> round(cellmean, 2)
  50-59 60-69 70-79 80-89  90+
F  2.62  2.91  3.22  3.94 4.98
M  2.83  3.22  3.91  4.65 6.00
```

Notice that the male/female difference in FLC varies with age, 2.6 versus 2.8 at age 50–59 and 5 versus 6 at age 90. The data does not fit a simple additive model; there are "interactions" to use statistical parlance. An excess of free light chain is thought to be at least partly a reflection of immune senescence, and due to our hormonal backgrounds men and women simply do not age in quite the same way.

4 Population averages

The question of how to test for a main effect in the presence of interaction is an old one. At one time this author considered the phrase "main effect in the presence of interaction" to be an oxymoron, but long experience with clinical data sets has led me to the opposite conclusion. Real data always has interactions. The treatment effect of a drug will not be exactly the same for old and young, thin and obese, physically active and sedentary, etc. Explicit recognition of this is an underlying rationale of the current drive towards "personalized medicine", though that buzzword often focuses only on genetic differences. Any given data set may often be too small to explore these variations and our statistical models will of necessity smooth over the complexity, but interactions are nevertheless still present.

Consider the data shown in figure **??** below, which shows a particular laboratory test value by age and sex. We see that the sex effect varies by age. Given this, what could be meant by a "main effect" of sex? One sensible approach is to select a fixed *population* for the ages, and then compute the average sex effect over that population. Indeed this is precisely what many computations do behind the scenes, e.g. the "type 3" estimates found in linear models.

There are three essential components to the calculation: a reference population for the confounders, a summary measure of interest, and a computational algorithm. To understand how linear models methods may (or may not) extend to the proportional hazards model it is useful consider all three facets; each is revealing.

Four possible choices for a target population of ages are given below.

1. Empirical: the age distribution of the sample at hand, also called the data distribution. In our sample this would be the age distribution of all 7874 subjects, ignoring sex.
2. SAS: a uniform distribution is assumed over all categorical adjusters, and the data distribution for continuous ones.
3. External reference: a fixed external population, e.g. the age distribution of the US 2010 census.
4. MVUE: minimum variance unbiased; the implicit population corresponding to a multivariate least squares fit.

Method 3 is common in epidemiology, method 1 is found in traditional survey sampling and in other common cases as we will see below. The type 3 estimates of SAS correspond to population 2. If there an interaction between two categorical variables x1 and x2, then the uniform distribution is taken to be over all combinations formed by the pair, and similarly for higher order interactions.

5 Linear models and populations

If we ignore the age effect, then everyone agrees on the best estimate of mean FLC: the simple average of FLC values within each sex. The male-female difference is estimated as the difference of these means. This is what is obtained from a simple linear regression of FLC on sex. Once we step beyond this and adjust for age, the relevant linear models can be looked at in several ways; we will explore three of them below: contrasts, case weights, and nesting. This "all roads lead to Rome" property of linear models is one of their fascinating aspects, at least mathematically.

5.1 Case weights

How do we form a single number summary of "the effect of sex on FLC"? Here are four common choices.

1. Unadjusted. The mean for males minus the mean for females. The major problem with this is that a difference in age distributions will bias the result. Looking at figure 1 imagine that this were two treatments A and B rather than male/female, and that the upper one had been given to predominantly 50-65 year olds and the lower predominantly to subjects over 80. An unadjusted difference would actually reverse the true ordering of the curves.
2. Population adjusted. An average difference between the curves, weighted by age. Three common weightings are
   (a) External reference. It is common practice in epidemiology to use an external population as the reference age distribution, for instance the US 2000 census distribution. This aids in comparing results between studies.

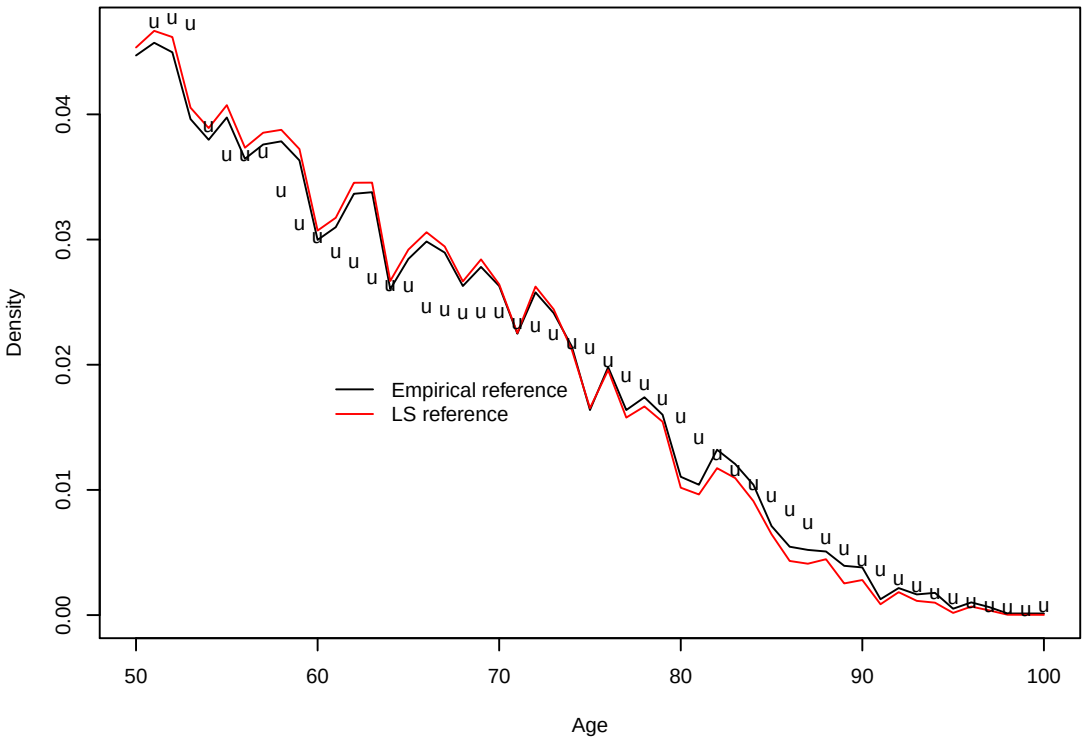

Figure 2: Three possible adjusting populations for the FLC data set, a empirical reference in black, least squares based one in red, and the US 2000 reference population as 'u'.

(b) Empirical population. The overall population structure of the observed data.

(c) Least squares. The population structure that minimizes the variance of the estimated female-male difference.

The principle idea behind case weights is to reweight the data such that confounders become balanced, i.e., ages are balanced when examining the sex effect and sex is balanced when examining age. Any fitted least squares estimate can be rewritten as a weighted sum of the data points with weight matrix $W = (X'X)^{-1}X'$. W has p rows, one per coefficient, each row is the weight vector for the corresponding element of $\hat{\beta}$. So we can backtrack and see what population assumption was underneath any given fit by looking at the weights for the relevant coefficient(s). Consider the two fits below. In both the second coefficient is an estimate of the overall difference in FLC values between the sexes. (The relationship in figure 1 is clearly curved so we have foregone the use of a simple linear term for age; there is no point in fitting an obviously incorrect model.) Since β_2 is a contrast the underlying weight vectors have negative values for the females and positive for the males.

```
> us2000 <- rowSums(uspop2[51:101,,'2000'])
> fit1 <- lm(flc ~ sex, flchain, x=TRUE)
> fit2 <- lm(flc ~ sex + ns(age,4), flchain, x=TRUE)
> c(fit1$coef[2], fit2$coef[2])
     sexM      sexM
0.2702554 0.3926299
> wt1  <- solve(t(fit1$x)%*%fit1$x, t(fit1$x))[2,] # unadjusted
> wt2 <-  solve(t(fit2$x)%*%fit2$x, t(fit2$x))[2,] # age-adjusted
> table(wt1, flchain$sex)
wt1                          F    M
  -0.000229885057471264 4350    0
  0.000283768444948922     0 3524
```

To reconstruct the implied population density, one can use the density function with `wt1` or `wt2` as the case weights. Examination of `wt1` immediately shows that the values are $-1/n_f$ for females and $1/n_m$ for males where n_f and n_m are number of males and females, respectively. The linear model `fit1` is the simple difference in male and female means; the implied population structure for males and females is the unweighted density of each.

Because this data set is very large and age is coded in years we can get a density estimate for fit2 by simple counting. The result is coded below and shown in figure 2. The empirical reference and least squares reference are nearly identical. This is not a surprise. Least squares fits produce minimum variance unbiased estimates (MVUE), and the variance of a weighted average is minimized by using weights proportional to the sample size, thus the MVUE estimate will give highest weights to those ages with a lot of people. The weights are not *exactly* proportional to sample size for each age. As we all know, for a given sample size n a study comparing two groups will have the most power with equal allocation between the groups. Because the M/F ratio is more unbalanced at the right edge of the age distribution the MVUE estimate gives just a little less weight there, but the difference between it and the overall data set population will be slight for all but those pathological cases where there is minimal overlap between M/F age

distributions. (And in that case the entire discussion about what "adjustment" can or should mean is much more difficult.)

```
> us2000 <- rowSums(uspop2[51:101,,'2000'])
> tab0 <- table(flchain$age)
> tab2 <- tapply(abs(wt2), flchain$age, sum)
> matplot(50:100, cbind(tab0/sum(tab0), tab2/sum(tab2)),
          type='l', lty=1,
          xlab="Age", ylab="Density")
> us2000 <- rowSums(uspop2[51:101,,'2000'])
> matpoints(50:100, us2000/sum(us2000), pch='u')
> legend(60, .02, c("Empirical reference", "LS reference"),
         lty=1, col=1:2, bty='n')
```

The LS calculation does a population adjustment automatically for us behind the scenes via the matrix algebra of linear models. If we try to apply population reference adjustment directly a problem immediately arises: in the US reference 0.13% of the population is aged 95 years, and our sample has no 95 year old males; it is not possible to re weight the sample so as to exactly match the US population reference. This occurs in any data set that is divided into small strata. The traditional epidemiology approach to this is to use wider age intervals of 5 or 10 years. Weights are chosen for each age/sex strata such that the sum of weights for females = sum of weights for males within each age group (balance), and the total sum of weights in an age group is equal to the reference population. The next section goes into this further. An increasingly popular approach for producing results that are standardized to the empirical reference population (i.e. the data distribution) is to use a smoothed age effect, obtained through inverse probability weights which are based on logistic regression, e.g. in the causal models literature and propensity score literature. This approach is illustrated in a vignette on adjusted survival curves which is also in the survival package.

5.2 Categorical predictors and contrasts

When the adjusting variable or variables are categorical — a factor in R or a class variable in SAS — then two more aspects come into play. The first is that any estimate of interest can be written in terms of the cell means. Formally, the cell means are a *sufficient statistic* for the data. For our data set and using the categorized variable `age2` let θ_{ij} parameterize these means.

	50–59	60–69	70-79	80-89	90+
Female	θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}
Male	θ_{21}	θ_{22}	θ_{23}	θ_{24}	θ_{25}

For a design with three factors we will have θ_{ijk}, etc. Because it is a sufficient statistic, any estimate or contrast of interest can be written as a weighted sum of the θs. Formulas for the resulting estimates along with their variances and tests were worked out by Yates in 1934 [14] and are often referred to as a Yates weighted means estimates. For higher order designs the computations can be rearranged in a form that is manageable on a desk calculator, and this is in fact the primary point of that paper. (Interestingly, his computational process turns out to be closely related to the fast Fourier transform.)

The second facet of categorical variables is that another adjustment is added to the list of common estimates:

1. Unadjusted
2. Population adjusted
   (a) External reference
   (b) Empirical (data set) reference
   (c) Least squares
   (d) Uniform. A population in which each combination of the factors has the same frequency of occurrence.

The uniform population plays a special role in the case of designed experiments, where equal allocation corresponds to the optimal study design. The Yates estimates are particularly simple in this case. For a hypothetical population with equal numbers in each age category the estimated average FLC for females turns out to be $\mu_f = \sum_j \theta_{1j}/5$ and the male - female contrast is $\sum_j(\theta_{2j} - \theta_{1j})/5$. We will refer to these as the "Yates" estimates and contrast for an effect. Conversely, the estimated age effects, treating sex as a confounding effect and assuming an equal distribution of females and males as the reference population, gives an estimated average FLC for the 60-69 year olds of $\mu_{60-69} = (\theta_{12} + \theta_{22})/2$, and etc for the other age groups.

We can obtain the building blocks for Yates estimates by using the interaction function and omitting the intercept.

```
> yatesfit <- lm(flc ~ interaction(sex, age2) -1, data=flchain)
> theta <- matrix(coef(yatesfit), nrow=2)
> dimnames(theta) <- dimnames(counts)
> round(theta,2)
   age2
sex 50-59 60-69 70-79 80-89  90+
  F  2.62  2.91  3.22  3.94 4.98
  M  2.83  3.22  3.91  4.65 6.00
```

For a linear model fit, any particular weighted average of the coefficients along with its variance and the corresponding sums of squares can be computed using the `contrast` function given below. Let C be a contrast matrix with k rows, each containing one column per coefficient. Then $C\theta$ is a vector of length k containing the weighted averages and $V = \hat{\sigma}^2 C(X'X)^{-1}C'$ is its variance matrix. The sums of squares is the increase in the sum of squared residuals if the fit were restricted to the subspace $C\theta = 0$. Formulas are from chapter 5 of Searle [8]. Some authors reserve the word *contrast* for the case where each row of C sums to zero and use *estimate* for all others; I am being less restrictive since the same computation serves for both.

```
> qform <- function(beta, var) # quadratic form b' (V-inverse) b
      sum(beta * solve(var, beta))
> contrast <- function(cmat, fit) {
      varmat <- vcov(fit)
      if (class(fit) == "lm") sigma2 <- summary(fit)$sigma^2
```

	estimate	sd	SS
Unadjusted	0.270	0.04157	142.2
MVUE: continuous age	0.393	0.04005	295.9
MVUE: categorical age	0.383	0.04019	282.5
Empirical (data) reference	0.392	0.04020	294.0
US200 reference	0.404	0.04065	306.0
Uniform (Yates)	0.590	0.09205	127.0

Table 1: Estimates of the male-female difference along with their standard errors. The last 4 rows are based on categorized age.

```
      else sigma2 <- 1   # for the Cox model case

      beta <- coef(fit)
      if (!is.matrix(cmat)) cmat <- matrix(cmat, nrow=1)
      if (ncol(cmat) != length(beta)) stop("wrong dimension for contrast")

      estimate <- drop(cmat %*% beta)  #vector of contrasts
      ss <- qform(estimate, cmat %*% varmat %*% t(cmat)) *sigma2
      list(estimate=estimate, ss=ss, var=drop(cmat %*% varmat %*% t(cmat)))
      }
> yates.sex <- matrix(0, 2, 10)
> yates.sex[1, c(1,3,5,7,9)] <-  1/5    #females
> yates.sex[2, c(2,4,6,8,10)] <- 1/5    #males
> contrast(yates.sex, yatesfit)$estimate  # the estimated "average" FLC for F/M
[1] 3.532048 4.121774
> contrast(yates.sex[2,]-yates.sex[,1], yatesfit) # male - female contrast
$estimate
[1] 0.5897261

$ss
[1] 126.962

$var
[1] 0.008472878
```

Table 2 shows all of the estimates of the male/female difference we have considered so far along with their standard errors. Because it gives a much larger weight to the 90+ age group than any of the other estimates, and that group has the largest M-F difference, the projected difference for a uniform population (Yates estimate) yields the largest contrast. It pays a large price for this in terms of standard error, however, and is over twice the value of the other approaches. As stated earlier, any least squares parameter estimate can be written as a weighted sum of the y values. Weighted averages have minimal variance when all of the weights are close to 1. The unadjusted estimate adheres to this precisely and the data-reference and MVUE stay as close as possible to constant weights, subject to balancing the population. The Yates estimate, by

		50–59	60–69	70–79	80–89	90+
Female	Unadjusted	1.00	1.00	1.00	1.00	1.00
	Min var	1.08	1.08	0.94	0.69	0.50
	Empirical	1.06	1.06	0.94	0.80	0.71
	Yates	0.53	0.72	0.92	1.90	10.74
Male	Unadjusted	1.00	1.00	1.00	1.00	1.00
	Min var	0.96	0.96	1.07	1.27	1.43
	Empirical	0.94	0.93	1.08	1.46	2.02
	Yates	0.47	0.63	1.05	3.49	30.64

Table 2: Observation weights for each data point corresponding to four basic approaches. All weights are normed so as to have an average value of 1.

treating every cell equally, implicitly gives much larger weights to the oldest ages. Table 2 shows the effective observation weights used for each of the age categories.

```
> casewt <- array(1, dim=c(2,5,4)) # case weights by sex, age group, estimator
> csum <- colSums(counts)
> casewt[,,2] <- counts[2:1,] / rep(csum, each=2)
> casewt[,,3] <- rep(csum, each=2)/counts
> casewt[,,4] <- 1/counts
> #renorm each so that the mean weight is 1
> for (i in 1:4) {
      for (j in 1:2) {
          meanwt <- sum(casewt[j,,i]*counts[j,])/ sum(counts[j,])
          casewt[j,,i] <- casewt[j,,i]/ meanwt
      }
  }
```

Looking at table 2 notice the per observation weights for the ≥ 90 age group, which is the one with the greatest female/male imbalance in the population. For all but the unbalanced estimate (which ignores age) the males are given a weight that is approximately 3 times that for females in order to re balance the shortage of males in that category. However, the absolute values of the weights differ considerably.

5.3 Different codings

Because the cell means are a sufficient statistic, all of the estimates based on categorical age can be written in terms of the cell means $\hat{\theta}$. The Yates contrast is the simplest to write down:

	50–59	60–69	70–79	80–89	90+
Female	-1/5	-1/5	-1/5	-1/5	-1/5
Male	1/5	1/5	1/5	1/5	1/5

For the data set weighting the values of 1/5 are replaced by n_{+j}/n_{++}, the overall frequency of each age group, where a + in the subscript stands for addition over that subscript in the table of

counts. The US population weights use the population frequency of each age group. The MVUE contrast has weights of $w_j/\sum w_j$ where $w_j = 1/(1/n_{1j} + 1/n_{2j})$, which are admittedly not very intuitive.

	50–59	60–69	70–79	80–89	90+
Female	-0.410	-0.303	-0.205	-0.073	-0.009
Male	0.410	0.303	0.205	0.073	0.009

In the alternate model `y ~sex + age2` the MVUE contrast is much simpler, namely (0, 1, 0,0,0,0,0), and can be read directly off the printout as $\beta/se(\beta)$. The computer's calculation of $(X'X)^{-1}$ has derived the "complex" MVUE weights for us without needing to lift a pencil. The Yates contrast, however, cannot be created from the coefficients of the simpler model at all. This observation holds in general: a contrast that is simple to write down in one coding may appear complicated in another, or not even be possible.

The usual and more familiar coding for a two way model is

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{1}$$

What do the Yates' estimates look like in this form? Let e_i be the Yates estimate for row i and k the number of columns in the two way table of θ values. Then

$$\begin{aligned} e_i &= (1/k) \sum_{j=1}^{k} \theta_{ij} \\ &= \mu + \alpha_i + \sum_j (\beta_j + \gamma_{ij})/k \end{aligned}$$

and the Yates test for row effect is

$$\begin{aligned} 0 &= e_i - e_{i'} \quad \forall i, i' \\ &= (\alpha_i - \alpha_{i'}) + (1/k) \sum_j (\gamma_{ij} - \gamma_{i'j}) \end{aligned} \tag{2}$$

Equation (1) is over determined and all computer programs add constraints in order to guarantee a unique solution. However those constraints are applied, however, equation (2) holds. The default in R is treatment contrasts, which use the first level of any factor as a reference level. Under this constraint the reference coefficients are set to zero, i.e., all coefficients of equations (1) and (2) above where $i = 1$ or $j = 1$. We have been computing the male - female contrast, corresponding to $i = 2$ and $i' = 1$ in equation (2), and the Yates contrast for sex becomes $\alpha_2 + 1/5(\gamma_{22} + \gamma_{23} + \gamma_{24} + \gamma_{25})$. The code below verifies that this contrast plus the usual R fit replicates the results in table 1.

```
> fit3 <- lm(flc ~ sex * age2, flchain)
> coef(fit3)
   (Intercept)           sexM      age260-69      age270-79
    2.61554420     0.21062295     0.29154970     0.60551059
     age280-89        age290+ sexM:age260-69 sexM:age270-79
    1.32247105     2.36298666     0.09772871     0.48323840
sexM:age280-89   sexM:age290+
    0.50065883     0.81388966
```

```
> contrast(c(0,1, 0,0,0,0, .2,.2,.2,.2), fit3) #Yates
$estimate
[1] 0.5897261

$ss
[1] 126.962

$var
[1] 0.008472878
```

The usual constraint is SAS is to use the last level of any class variable as the reference group, i.e., all coefficients with $i = 2$ or $j = 5$ in equations (1) and (2) are set to zero.

```
> options(contrasts=c("contr.SAS", "contr.poly"))
> sfit1 <- lm(flc ~ sex, flchain)
> sfit2 <- lm(flc ~ sex + age2, flchain)
> sfit3 <- lm(flc ~ sex * age2, flchain)
> contrast(c(0,-1, 0,0,0,0, -.2,-.2,-.2,-.2), sfit3) # Yates for SAS coding
$estimate
[1] 0.5897261

$ss
[1] 126.962

$var
[1] 0.008472878
```

The appendix contains SAS code and output for the three models `sfit1, sfit2` and `sfit3` above. The `E3` option was added to the SAS model statements, which causes a symbolic form of the contrasts that were used for "type III" results to be included in the printout. Look down the column labeled "SEX" and you will see exactly the coefficients used just above, after a bit of SAS to English translation.

- The SAS printout is labeled per equation (1), so L1= column 1 of the full X matrix = intercept. L2 = column 2 = females, L3 = column 3 = males, L4= column 4 = age 50–59, etc.
- In the symbolic printout they act as though sum constraints were in force: the last column of age is labeled with a symbolic value that would cause the age coefficients to sum to zero. However, in actuality these coefficients are set to zero. The table of parameter estimates at the end of the printout reveals this; forced zeros have a blank for their standard error.
- When calculating the contrast one can of course skip over the zero coefficients, and the R functions do not include them in the coefficient vector. Remove all of these aliased rows from the SAS symbolic printout to get the actual contrast that is used; this will agree with my notation.
- The SAS printout corresponds to a female-male contrast and I have been using male-female for illustration. This changes the signs of the contrast coefficients but not the result.

The `estimate` statement in the SAS code required that all of the coefficients be listed, even the aliased ones (someone more proficient in SAS may know a way to avoid this and enter only the non-aliased values.)

So, how do we actually compute the Yates contrast in a computer program? We will take it as a give that no one wants to memorize contrast formulas. Appendix A describes three algorithms for the computation.

One of these three (NSTT) is completely unreliable, but is included because it is so often found in code. If one uses the sum constraints commonly found in textbooks, which corresponds to the `contr.sum` constraint in R and to `effect` constraints in SAS, and there are no missing cells, then the last term in equation (2) is zero and the simple contrast $\alpha_i = 0$ will be equal to the Yates contrast for sex. I often see this method recommended on R help in response to the question of "how to obtain type III", computed either by use of the `drop1` command or the `Anova` function found within the car package, but said advice almost never mentions the need for this particular non-default setting of the contrasts option[2]. When applied to other codings the results of this procedure can be surprising.

```
> options(contrasts = c("contr.treatment", "contr.poly")) #R default
> fit3a <- lm(flc ~ sex * age2, flchain)
> options(contrasts = c("contr.SAS", "contr.poly"))
> fit3b <- lm(flc~ sex * age2, flchain)
> options(contrasts=c("contr.sum", "contr.poly"))
> fit3c <- lm(flc ~ sex * age2, flchain)
> #
> nstt <- c(0,1, rep(0,8))  #test only the sex coef = the NSTT method
> temp <- rbind(unlist(contrast(nstt, fit3a)),
                unlist(contrast(nstt, fit3b)),
                unlist(contrast(nstt, fit3c)))[,1:2]
> dimnames(temp) <- list(c("R", "SAS", "sum"), c("effect", "SS"))
> print(temp)
       effect        SS
R    0.210623  34.94679
SAS -1.024513  18.80244
sum -0.294863 126.96199
> #
> drop1(fit3a, .~.)
Single term deletions

Model:
flc ~ sex * age2
         Df Sum of Sq   RSS    AIC
<none>                24325 8901.3
sex       1     34.95 24360 8910.6
age2      4   1020.55 25345 9216.9
sex:age2  4     87.22 24412 8921.5
```

[2]The Companion to Applied Regression (car) package is designed to be used with the book of the same name by John Fox, and the book does clarify the need for sum constraints.

For the case of a two level effect such as sex, the NSTT contrast under the default R coding is a comparison of males to females in the first age group **only**, and under the default SAS coding it is a comparison of males to females within the **last** age group. Due to this easy creation of a test statistic which has no relation to the global comparison one expects from the "type 3" label the acronym *not safe type three*(NSTT) was chosen, "not SAS" and "nonsense" are alternate mnemonics.

5.4 Sums of squares and projections

The most classic exposition of least squares is as a set of projections, each on to a smaller space. Computationally we represent this as a series of model fits, each fit summarized by the change from the prior fit in terms of residual sum of squares.

```
> options(show.signif.stars = FALSE) #exhibit intelligence
> sfit0  <- lm(flc ~ 1, flchain)
> sfit1b <- lm(flc ~ age2, flchain)
> anova(sfit0, sfit1b, sfit2, sfit3)
Analysis of Variance Table

Model 1: flc ~ 1
Model 2: flc ~ age2
Model 3: flc ~ sex + age2
Model 4: flc ~ sex * age2
  Res.Df   RSS Df Sum of Sq        F    Pr(>F)
1   7873 26624
2   7869 24694  4   1929.64 155.9598 < 2.2e-16
3   7868 24412  1    282.49  91.3284 < 2.2e-16
4   7864 24325  4     87.22   7.0493 1.163e-05
```

The second row is a test for the age effect. The third row of the above table summarizes the improvement in fit for the model with sex + age2 over the model with just age2, a test of "sex, adjusted for age". This test is completely identical to the minimum variance contrast, and is in fact the way in which that SS is normally obtained. The test for a sex effect, unadjusted for age, is identical to an anova table that compares the intercept-only fit to one with sex, i.e., the second line from a call to `anova(sfit0, sfit1)`.

The anova table for a nested sequence of models A, $A + B$, $A + B + C$, ... has a simple interpretation, outside of contrasts or populations, as an improvement in fit. Did the variable(s) B add significantly to the goodness of fit for a model with just A, was C an important addition to a model that already includes A and B? The assessment of improvement is based on the likelihood ratio test (LRT), and extends naturally to all other models based on likelihoods. The tests based on a target population (external, data population, or Yates) do not fit naturally into this approach, however.

5.5 What is SAS type 3?

We are now in a position to fully describe the SAS sums of squares.

- Type 1 is the output of the ANOVA table, where terms are entered in the order specified in the model.
- Type 2 is the result of a two stage process
  1. Order the terms by level: 0= intercept, 1= main effects, 2= 2 way interactions, ...
  2. For terms of level k, print the MVUE contrast from a model that includes all terms of levels $0 - k$. Each of these will be equivalent to the corresponding line of a sequential ANOVA table where the term in question was entered as the last one of its level.
- Type 3 and 4 are also a 2 stage process
  1. Segregate the terms into those for which a Yates contrast can be formed versus those for which it can not. The second group includes the intercept, any continuous variables, and any factor (class) variables that do not participate in interactions with other class variables.
  2. For variables in the first group compute Yates contrasts. For those in the second group compute the type 2 results.

SAS has two different algorithms for computing the Yates contrast, which correspond to the `ATT` and `STT` options of the `yates` function. SAS describes the two contrast algorithms in their document "The four types of estimable functions" [7], one of which defines type 3 and the other type 4. I found it very challenging to recreate their algorithm from this document. Historical knowledge of the underlying linear model algorithms used by SAS is a useful and almost necessary adjunct, as many of the steps in the document are side effects of their calculation.

When there are missing cells, then it is not possible to compute a contrast that corresponds to a uniform distribution over the cells, and thus the standard Yates contrast is also not defined. The SAS type 3 and 4 algorithms still produce a value, however. What exactly this result "means" and whether it is a good idea has been the subject of lengthy debates which I will not explore here. Sometimes the type 3 and type 4 algorithms will agree but often do not when there are missing cells, which further muddies the waters.

Thus we have 3 different tests: the MVUE comparison which will be close but not exactly equal to the data set population, Yates comparisons which correspond to a uniform reference population, and the SAS type 3 (STT) which prints out a chimeric blend of uniform population weighting for those factor variables that participate in interactions and the MVUE weighting for all the other terms.

5.6 Which estimate is best?

Deciding which estimate is the best is complicated. Unfortunately a lot of statistical textbooks emphasize the peculiar situation of balanced data with exactly the same number of subjects in each cell. Such data is *extremely* peculiar if you work in medicine; in 30 years work and several hundred studies I have seen 2 instances. In this peculiar case the unadjusted, MVUE, empirical reference and Yates populations are all correspond to a uniform population and so give identical results. No thinking about which estimate is best is required. This has led many to avoid the above question, instead pining for that distant Eden where the meaning of "row effect" is perfectly unambiguous. But we are faced with real data and need to make a choice.

The question has long been debated in depth by wiser heads than mine. In a companion paper to his presentation at the joint statistical meetings in 1992, Macnaughton [5] lists 54 references to the topic between 1952 and 1991. Several discussion points recur:

1. Many take the sequential ANOVA table as primary, i.e., a set of nested models along with likelihood ratio tests (LRT), and decry all comparisons of "main effects in the presence of interaction." Population weightings other than the LS one do not fit nicely into the nested framework.

2. Others are distressed by the fact that the MVUE adjusting population is data dependent, so that one is "never sure exactly what hypothesis being tested".

3. A few look at the contrast coefficients themselves, with a preference for simple patterns since they "are interpretable".

4. No one approach works for all problems. Any author who proposes a uniform rule is quickly presented with counterexamples.

Those in group 1 argue strongly against the Yates weighting and those in group 2 argue for the Yates contrast. Group 3 is somewhat inexplicable to me since any change in the choice of constraint type will change all the patterns. I fear that an opening phrase from the 1986 overview/review of Herr [3] is still apropos, "In an attempt to understand how we have arrived at our present state of ignorance . . . ".

There are some cases where the Yates approach is clearly sensible, for instance a designed experiment which has become unbalanced due to a failed assay or other misadventure that has caused a few data points to be missing. There are cases such as the FLC data where the Yates contrast makes little sense at all — the hypothetical population with equal numbers of 50 and 90 year olds is one that will never be seen— so it is rather like speculating on the the potential covariate effect in dryads and centaurs. The most raucous debate has circled around the case of testing for a treatment effect in the presence of multiple enrolling centers. Do we give each patient equal weight (MVUE) or each center equal weight (Yates). A tongue-in-cheek but nevertheless excellent commentary on the subject is given by the old curmudgeon, aka Guernsey McPearson [9, 10]. A modern summary with focus on the clinical trials arena is found in chapter 14 of the textbook by Senn [11]

I have found two papers particularly useful in thinking about this. Senn **??** points out the strong parallels between tests for main effects when there may be interactions and meta analyses, cross connecting these two approaches is illuminating. A classic reference is the 1978 paper by Aitkin [1]. This was read before the Royal Statistical Society and includes remarks by 10 discussants forming a who's who of statistical theory (F Yates, J Nelder, DR Cox, DF Andrews, KR Gabriel, . . .). The summary of the paper states that "It is shown that a standard method of analysis used in many ANOVA programs, equivalent to Yates method of weighted squares of means, may lead to inappropriate models"; the paper goes on to carefully show why no one method can work in all cases. Despite the long tradition among RSS discussants of first congratulating the speaker and then skewering every one their conclusions, not one defense of the always-Yates approach is raised! This includes the discussion by Yates himself, who protests that his original paper advocated the proposed approach with reservations, it's primary advantage being that the computations could be performed on a desk calculator.

I have two primary problems with the SAS type 3 approach. The first and greatest is that their documentation recommends the method with no reference to this substantial and sophisticated literature discussing strengths and weaknesses of the Yates contrast. This represents a level of narcissism which is completely unprofessional. The second is that their documentation explains the method is a way that is almost impenetrably opaque. If this is the only documentation one has, there will not be 1 statistician in 20 who would be able to explain the actual biological hypothesis which is being addressed by a type 3 test.

6 Cox models

6.1 Tests and contrasts

Adapting the Yates test to a Cox model is problematic from the start. First, what do we mean by a "balanced population"? In survival data, the variance of the hazard ratio for each particular sex/age combination is proportional to the number of deaths in that cell rather than the number of subjects. Carrying this forward to the canonical problem of adjusting a treatment effect for enrolling center, does this lead to equal numbers of subjects or equal numbers of events? Two centers might have equal numbers of patients but different number of events because one initiated the study at a later time (less follow up per subject), or it might have the same follow up time but a lower death rate. Should we reweight in one case (which one), both, or neither? The second issue is that the per-cell hazard ratio estimates are no longer a minimally sufficient statistic, so underlying arguments about a reference population no longer directly translate into a contrast of the parameters. A third but more minor issue is that the three common forms of the test statistic — Wald, score, and LRT — are identical in a linear model but not for the Cox model, so which should we choose?

To start, take a look at the overall data and compute the relative death rates for each age/sex cell.

```
> options(contrasts= c("contr.treatment", "contr.poly")) # R default
> cfit0 <- coxph(Surv(futime, death) ~ interaction(sex, age2), flchain)
> cmean <- matrix(c(0, coef(cfit0)), nrow=2)
> cmean <- rbind(cmean, cmean[2,] - cmean[1,])
> dimnames(cmean) <- list(c("F", "M", "M/F ratio"), dimnames(counts)[[2]])
> signif(exp(cmean),3)
          50-59 60-69 70-79 80-89   90+
F          1.00  2.45  7.28 21.90 69.30
M          1.46  3.88 11.10 27.00 81.60
M/F ratio  1.46  1.58  1.53  1.23  1.18
```

Since the Cox model is a relative risk model all of the death rates are relative to one of the cells, in this case the 50–59 year old females has been arbitrarily chosen as the reference cell and so has a defined rate of 1.00. Death rates rise dramatically with age for both males and females (no surprise), with males always slightly ahead in the race to a coffin. The size of the disadvantage for males decreases in the last 2 decades, however.

The possible ways to adjust for age in comparing the two sexes are

1. The likelihood ratio test. This is analogous to the sequential ANOVA table in a linear model, and has the strongest theoretical justification.

2. A stratified Cox model, with age group as the stratification factor. This gives a more general and rigorous adjustment for age. Stratification on institution is a common approach in clinical trials.

3. The Wald or score test for the sex coefficient, in a model that adjusts for age. This is analogous to Wald tests in the linear model, and is asymptotically equivalent the the LRT.

4. The test from a reweighted model, using case weights. Results using this approach have been central to causal model literature, particularly adjustment for covariate imbalances in observational studies. (Also known as *marginal structural models*). Adjustment to a uniform population is also possible.

5. A Yates-like contrast in the Cox model coefficients.

    - A reliable algorithm such as cell means coding.
    - Unreliable approach such as the NSTT

I have listed these in order from the most to the least available justification, both in terms of practical experience and available theory. The two standard models are for sex alone, and sex after age. Likelihood ratio tests for these models are the natural analog to anova tables for the linear model, and are produced by the same R command. Here are results for the first three, along with the unadjusted model that contains sex only.

```
> options(contrasts=c("contr.SAS", "contr.poly"))
> cfit1 <- coxph(Surv(futime, death) ~ sex, flchain)
> cfit2 <- coxph(Surv(futime, death) ~ age2 + sex, flchain)
> cfit3 <- coxph(Surv(futime, death) ~ sex + strata(age2), flchain)
> # Unadjusted
> summary(cfit1)
Call:
coxph(formula = Surv(futime, death) ~ sex, data = flchain)

  n= 7874, number of events= 2169

         coef exp(coef) se(coef)      z Pr(>|z|)
sexF -0.08413   0.91932  0.04307 -1.953   0.0508

     exp(coef) exp(-coef) lower .95 upper .95
sexF    0.9193      1.088    0.8449         1

Concordance= 0.509  (se = 0.005 )
Rsquare= 0   (max possible= 0.992 )
Likelihood ratio test= 3.81  on 1 df,   p=0.05105
Wald test            = 3.82  on 1 df,   p=0.05077
Score (logrank) test = 3.82  on 1 df,   p=0.05071
```

```
> #
> # LRT
> anova(cfit2)
Analysis of Deviance Table
 Cox model: response is Surv(futime, death)
Terms added sequentially (first to last)

     loglik    Chisq Df Pr(>|Chi|)
NULL -18868
age2 -17724 2288.464  4  < 2.2e-16
sex  -17690   69.059  1  < 2.2e-16
> #
> # Stratified
> anova(cfit3)
Analysis of Deviance Table
 Cox model: response is Surv(futime, death)
Terms added sequentially (first to last)

     loglik  Chisq Df Pr(>|Chi|)
NULL -14668
sex  -14633 69.648  1  < 2.2e-16
> summary(cfit3)
Call:
coxph(formula = Surv(futime, death) ~ sex + strata(age2), data = flchain)

  n= 7874, number of events= 2169

        coef exp(coef) se(coef)    z Pr(>|z|)
sexF -0.3679    0.6922   0.0438 -8.4   <2e-16

     exp(coef) exp(-coef) lower .95 upper .95
sexF    0.6922      1.445    0.6352    0.7542

Concordance= 0.55  (se = 0.012 )
Rsquare= 0.009   (max possible= 0.976 )
Likelihood ratio test= 69.65  on 1 df,   p=1.11e-16
Wald test            = 70.56  on 1 df,   p=0
Score (logrank) test = 71.29  on 1 df,   p=0
> #
> # Wald test
> signif(summary(cfit2)$coefficients, 3)
            coef exp(coef) se(coef)      z Pr(>|z|)
age250-59 -4.180    0.0153   0.1220 -34.30 0.00e+00
age260-69 -3.240    0.0392   0.1140 -28.40 0.00e+00
age270-79 -2.180    0.1140   0.1100 -19.80 0.00e+00
```

```
age280-89 -1.150    0.3160   0.1110 -10.40 0.00e+00
sexF      -0.366    0.6930   0.0438  -8.36 1.11e-16
> #
> anova(cfit1, cfit2)
Analysis of Deviance Table
 Cox model: response is  Surv(futime, death)
 Model 1: ~ sex
 Model 2: ~ age2 + sex
  loglik  Chisq Df P(>|Chi|)
1 -18866
2 -17690 2353.7  4 < 2.2e-16
```

Without adjustment for age the LRT for sex is only 3.8, and after adjustment for a it increases to 69.06. Since females are older, not adjusting for age almost completely erases the evidence of their actual survival advantage. Results of the LRT are unchanged if we change to any of the other possible codings for the factor variables (not shown). Adjusting for age group using a stratified model gives almost identical results to the sequential LRT, in this case. The Wald tests for sex are equal to $[\beta/se(\beta)]^2$ using the sex coefficient from the fits, 3.82 and 69.96 for the unadjusted and adjusted models, respectively. Unlike a linear model they are not exactly equal to the anova table results based on the log-likelihood, but tell the same story.

Now consider weighted models, with both empirical and uniform distributions as the target age distribution. The fits require use of a robust variance, since we are approaching it via a survey sampling computation. The tapply function creates a per-subject index into the case weight table created earlier.

```
> wtindx <- with(flchain, tapply(death, list(sex, age2)))
> cfitpop <- coxph(Surv(futime, death) ~ sex, flchain,
                   robust=TRUE, weight = (casewt[,,3])[wtindx])
> cfityates <- coxph(Surv(futime, death) ~ sex, flchain,
                   robust=TRUE, weight = (casewt[,,4])[wtindx])
> #
> # Glue it into a table for viewing
> #
> tfun <- function(fit, indx=1) {
      c(fit$coef[indx], sqrt(fit$var[indx,indx]))
    }
> coxp <- rbind(tfun(cfit1), tfun(cfit2,5), tfun(cfitpop), tfun(cfityates))
> dimnames(coxp) <- list(c("Unadjusted", "Additive", "Empirical Population",
                           "Uniform Population"),
                         c("Effect", "se(effect)"))
> signif(coxp,3)
                      Effect se(effect)
Unadjusted           -0.0841     0.0431
Additive             -0.3660     0.0438
Empirical Population -0.3130     0.0435
Uniform Population   -0.2060     0.0761
```

The population estimates based on reweighting lie somewhere between the unadjusted and the sequential results. We expect that balancing to the empirical population will give a solution that is similar to the age + sex model, in the same way that the close but not identical to the MVUE estimate in a linear model. Balancing to a hypothetical population with equal numbers in each age group yields a substantially smaller estimate of effect. since it gives large weights to the oldest age group, where in this data set the male/female difference is smallest.

Last, look at constructed contrasts from a cell means model. We can either fit this using the interaction, or apply the previous contrast matrix to the coefficients found above. Since the "intercept" of a Cox model is absorbed into the baseline hazard our contrast matrix will have one less column.

```
> cfit4 <- coxph(Surv(futime, death) ~ sex * age2, flchain)
> # Uniform population contrast
> ysex <- c(0,-1, 0,0,0,0, -.2,-.2,-.2,-.2) #Yates for sex, SAS coding
> contrast(ysex[-1], cfit4)
$estimate
[1] 0.3260088

$ss
[1] 28.09557

$var
[1] 0.003782865
> # Verify using cell means coding
> cfit4b <- coxph(Surv(futime, death) ~ interaction(sex, age2), flchain)
> temp <- matrix(c(0, coef(cfit4b)),2)  # the female 50-59 is reference
> diff(rowMeans(temp)) #direct estimate of the Yates
[1] -0.3260088
> #
> temp2 <- rbind(temp, temp[2,] - temp[1,])
> dimnames(temp2) <- list(c('female', 'male', 'difference'), levels(age2))
> round(temp2, 3)
            50-59  60-69  70-79  80-89    90+
female      0.000 -4.026 -3.047 -1.994 -1.105
male       -4.402 -3.505 -2.416 -1.314 -0.164
difference -4.402  0.520  0.630  0.680  0.941
> #
> #
> # NSTT contrast
> contrast(c(1,0,0,0,0,0,0,0,0,0), cfit4)
$estimate
[1] -0.1639796

$ss
[1] 0.4746563
```

```
$var
[1] 0.0566501
```

In the case of a two level covariate such as sex, the NSTT algorithm plus the SAS coding yields an estimate and test for a difference in sex for the *first* age group; the proper contrast is an average. Since it gives more weight to the larger ages, where the sex effect is smallest, the Yates-like contrast is smaller than the result from an additive model `cfit2`. Nevertheless, this contrast and the sequential test are more similar for the survival outcome than for the linear models. This is due to the fact that the variances of the individual hazards for each sex/age combination are proportional to the number of deaths in that cell rather than the number of subjects per cell. A table of the number of deaths is not as imbalanced as the table of subject counts, and so the Yates and MLE "populations" are not as far apart as they were for the linear regression. There are fewer subjects at the higher ages but they die more frequently.

Why is the Yates-like contrast so different than the result of creating a uniform age distribution using case weights followed by an MLE estimate? Again, the MLE estimate has death counts as the effective weights; the case-weighted uniform population has smaller weights for the youngest age group and that group also has the lowest death rate, resulting in lower influence for that group and an estimate shrunken towards the 90+ difference of 0.941. All told, for survival models adjustment to a uniform population is a slippery target.

6.2 SAS phreg results

Now for the main event: what does SAS do? First, for the simple case of an additive model the SAS results are identical to those shown above. The coefficients, variances and log-likelihoods for cfit2 are identical to the phreg output for an additive model, as found in the appendix. As would be expected from the linear models case, the "type III" results for the additive model are simply the Wald tests for the fit, repackaged with a new label.

Now look at the model that contains interactions. We originally surmised that a contrast calculation would be the most likely way in which the phreg code would implement type 3, as it is the easiest to integrate with existing code. Results are shown in the last SAS fit of the appendix. Comparing these results of the SAS printout labeled as "Type III Wald" to the contrasts calculated above shows that phreg is using the NSTT method. This is a bit of a shock. All of the SAS literature on type III emphasizes the care with which they form the calculation so as to always produce a Yates contrast (or in the case of missing cells a Yates-like one), and there was no hint in the documentation that phreg does anything different. As a double check direct contrast statements corresponding to the Yates and NSTT contrasts were added to the SAS code, and give confirmatory results. A further run which forced sum constraints by adding `'/effect'` to the SAS class statement (not shown) restored the correct Yates contrast, as expected. As a final check, look at the NSTT version of the LRT, which corresponds to simply dropping the sex column from the X matrix.

```
> xmat4 <- model.matrix(cfit4)
> cfit4b <- coxph(Surv(futime, death) ~ xmat4[,-1], flchain)
> anova(cfit4b, cfit4)
Analysis of Deviance Table
 Cox model: response is  Surv(futime, death)
```

```
 Model 1: ~ xmat4[, -1]
 Model 2: ~ sex * age2
  loglik  Chisq Df P(>|Chi|)
1 -17687
2 -17687 0.4607  1    0.4973
```

This agrees with the LR "type 3" test of the phreg printout.

6.3 Conclusion

Overall, both rebalanced estimates and coefficient contrasts are interesting exercises for the Cox model, but their actual utility is unclear. It is difficult to make a global optimality argument for either one, particularly in comparison to the sequential tests which have the entire weight of likelihood theory as a justification. Case reweighted estimates do play a key role when attempting to adjust for non-random treatment assignment, as found in the literature for causal analysis and marginal structural models; a topic and literature far too extensive and nuanced for discussion in this note.

No special role is apparent, at least to this author, for regular or even sporadic use of a Yates contrast in survival models. The addition of such a feature and label to the SAS phreg package is a statistical calamity, one that knowledgeable and conscientious statistical practitioners will likely have to fight for the rest of their careers. In the common case of a treatment comparison, adjusted for enrolling center, the default "type III" printout from phreg corresponds to a comparison of treatments within the last center; the only contribution of the remainder of the data set is to help define the baseline hazard function and the effect of any continuous adjusters that happen to be in the model. The quadruple whammy of a third rate implementation (the NSTT), defaults that lead to a useless and misleading result, no documentation of the actual computation that is being done, and irrational reverence for the type III label conspire to make this a particularly unfortunate event.

A Computing the Yates estimate

We will take it as a given that no one wants to memorize contrast formulas, and so we need a way to compute Yates contrasts automatically in a computer program. The most direct method is to encode the original fit in terms of the cell means, as has been done throughout this report. The Yates contrast is then simply an average of estimates across the appropriate margin. However, we normally will want to solve the linear or Cox model fit in a more standard coding and then compute the Yates contrast after the fact. Note that any population re norming requires estimates of the cell means, whether they were explicit parameters or not, i.e., the model fit must include interaction terms.

Here are three algorithms for this post-hoc computation. All of them depend, directly or

indirectly, on the breakdown found earlier in equation (1).

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon && (3)\\ &= \theta_{ij} + \epsilon && (4)\\ \theta_{ij} &= \mu + \alpha_i + \beta_j + \gamma_{ij} && (5)\\ & && (6) \end{aligned}$$

Equation (3) is the standard form from our linear models textbooks, equation (4) is the cell means form, and (5) is the result of matching them together. Using this equivalence a Yates test for row effects will be

$$\begin{aligned} 0 &= e_i - e_{i'} \quad \forall i, i' \\ &= (\alpha_i - \alpha_{i'}) + (1/k) \sum_j (\gamma_{ij} - \gamma_{i'j}) \end{aligned} \tag{7}$$

where the subscripts i and i' range over the rows and k is the number of columns.

To illustrate the methods we will use 3 small data sets defined below. All are unbalanced. The second data set removes the aD observation and so has a zero cell, the third removes the diagonal and has 3 missing cells.

```
> data1 <- data.frame(y = rep(1:6, length=20),
                      x1 = factor(letters[rep(1:3, length=20)]),
                      x2 = factor(LETTERS[rep(1:4, length=10)]),
                      x3 = 1:20)
> data1$x1[19] <- 'c'
> data1 <- data1[order(data1$x1, data1$x2),]
> row.names(data1) <- NULL
> with(data1, table(x1,x2))
   x2
x1  A B C D
  a 1 2 2 1
  b 2 2 1 2
  c 3 2 1 1
> # data2 -- single missing cell
> indx <- with(data1, x1=='a' & x2=='D')
> data2 <- data1[!indx,]
> #data3 -- missing the diagonal
> data3 <- data1[as.numeric(data1$x1) != as.numeric(data1$x2),]
```

A.1 NSTT method

The first calculation method is based on a simple observation. If we impose the standard sums constraint on equation (3) which is often found in textbooks (but nowhere else) of $\sum_i \alpha_i = \sum_j \beta_j = 0$, $\sum_i \gamma_{ij} = 0 \; \forall j$ and $\sum_j \gamma_{ij} = 0 \; \forall i$, then the last term in equation (7) is identically 0. Thus the Yates contrast corresponds exactly to a test of $\alpha = 0$. In R we can choose this coding

by using the `contr.sum` option. This approach has the appearance of simplicity: we can do an ordinary test for row effects within an interaction model. Here is R code that is often proposed for "type III" computation, which is based on the same process.

```
> options(contrasts=c("contr.sum", "contr.poly"))
> fit1 <- lm(y ~ x1*x2, data1)
> drop1(fit1, .~.)
Single term deletions

Model:
y ~ x1 * x2
       Df Sum of Sq    RSS      AIC
<none>               2.667 -16.2981
x1      2     9.232 11.899   9.6143
x2      3     5.844  8.511   0.9122
x1:x2   6    43.858 46.524  28.8848
```

The problem with this approach is that it depends critically on use of the sum constraints. If we apply the same code after fitting the data set under the more usual constraints a completely different value ensues.

```
> options(contrasts=c("contr.SAS", "contr.poly"))
> fit2 <- lm(y ~ x1*x2, data1)
> drop1(fit2, .~.)
Single term deletions

Model:
y ~ x1 * x2
       Df Sum of Sq    RSS     AIC
<none>               2.667 -16.298
x1      2    11.000 13.667  12.385
x2      3    21.333 24.000  21.646
x1:x2   6    43.858 46.524  28.885
> options(contrasts=c("contr.treatment", "contr.poly"))
> fit3 <- lm(y ~ x1*x2, data1)
> drop1(fit3, .~.)
Single term deletions

Model:
y ~ x1 * x2
       Df Sum of Sq    RSS     AIC
<none>               2.667 -16.298
x1      2    13.333 16.000  15.537
x2      3    13.500 16.167  13.744
x1:x2   6    43.858 46.524  28.885
```

Both common choices of contrasts give a different answer than contr.sum, and both are useless. I thus refer to this as the Not Safe Type Three (NSTT) algorithm, "not SAS type three" and "nonsense type three" are two other sensible expansions. This approach should NEVER be used in practice.

A.2 ATT

The key idea of the averaging approach (Averaged Type Three) is to directly evaluate equation (7). The first step of the computation is shown below

```
> X <- model.matrix(fit2)
> ux <- unique(X)
> ux
   (Intercept) x1a x1b x2A x2B x2C x1a:x2A x1b:x2A x1a:x2B
1            1   1   0   1   0   0       1       0       0
2            1   1   0   0   1   0       0       0       1
4            1   1   0   0   0   1       0       0       0
6            1   1   0   0   0   0       0       0       0
7            1   0   1   1   0   0       0       1       0
9            1   0   1   0   1   0       0       0       0
11           1   0   1   0   0   1       0       0       0
12           1   0   1   0   0   0       0       0       0
14           1   0   0   1   0   0       0       0       0
17           1   0   0   0   1   0       0       0       0
19           1   0   0   0   0   1       0       0       0
20           1   0   0   0   0   0       0       0       0
   x1b:x2B x1a:x2C x1b:x2C
1        0       0       0
2        0       0       0
4        0       1       0
6        0       0       0
7        0       0       0
9        1       0       0
11       0       0       1
12       0       0       0
14       0       0       0
17       0       0       0
19       0       0       0
20       0       0       0
> indx <- rep(1:3, c(4,4,4))
> effects <- t(rowsum(ux, indx)/4)  # turn sideways to fit the paper better
> effects
               1    2    3
(Intercept) 1.00 1.00 1.00
x1a         1.00 0.00 0.00
x1b         0.00 1.00 0.00
```

```
x2A         0.25 0.25 0.25
x2B         0.25 0.25 0.25
x2C         0.25 0.25 0.25
x1a:x2A     0.25 0.00 0.00
x1b:x2A     0.00 0.25 0.00
x1a:x2B     0.25 0.00 0.00
x1b:x2B     0.00 0.25 0.00
x1a:x2C     0.25 0.00 0.00
x1b:x2C     0.00 0.25 0.00
> yates <- effects[,-1] - effects[,1]
> yates
                2     3
(Intercept)  0.00  0.00
x1a         -1.00 -1.00
x1b          1.00  0.00
x2A          0.00  0.00
x2B          0.00  0.00
x2C          0.00  0.00
x1a:x2A     -0.25 -0.25
x1b:x2A      0.25  0.00
x1a:x2B     -0.25 -0.25
x1b:x2B      0.25  0.00
x1a:x2C     -0.25 -0.25
x1b:x2C      0.25  0.00
```

The data set ux has 12 rows, one for each of the 12 unique x*x2 combinations. Because data1 was sorted, the first 4 rows correspond to x=1, the next 4 to x=2 and the next to x=3 which is useful for illustration but has no impact on the computation. The average of rows 1-4 (column 1 of `effects` above) is the estimated average response for subjects with x1=a, assuming a uniform distribution over the 12 cells. Any two differences between the three effects is an equivalent basis for computing the Yates contrast.

We can verify that the resulting estimates correspond to a uniform target population by directly examining the case weights for the estimate. Each of them gives a total weight of 1/4 to each level of x2. Each element of $\beta\beta$ is a weighted average of the data, revealed by the rows of the matrix $(X'X)^{-1}X'$. The estimate are a weighted sum of the coefficients, so are also a weighted average of the y values.

```
> wt <- solve(t(X) %*% X, t(X)) # twelve rows (one per coef), n columns
> casewt <- t(effects) %*% wt   # case weights for the three "row efffects"
> for (i in 1:3) print(tapply(casewt[i,], data1$x2, sum))
   A    B    C    D
0.25 0.25 0.25 0.25
   A    B    C    D
0.25 0.25 0.25 0.25
   A    B    C    D
0.25 0.25 0.25 0.25
```

A.3 STT

The SAS type III method takes a different approach, based on a a dependency matrix D. Start by writing the X matrix for the problem using all of the parameters in equation (3). For our flc example this will have columns for intercept (1), sex (2), age group (5) and the age group by sex interaction (10) = 18 columns. Now define the lower triangular square matrix D such that

- If the ith column of X can be written as a linear combination of columns 1 through $i-1$, then row i of D contains that linear combination and $D_{ii} = 0$.
- If the ith column is not linearly dependent on earlier ones then $D_{ii} = 1$ and $D_{ij} = 0$ for all $j \neq i$.

Columns of D that correspond to linearly dependent columns of X will be identically zero and can be discarded (or not) at this point. The result of this operation replicates table 12.2 in the SAS reference [7] labeled "the form of estimable functions". To obtain the Yates contrasts for an effect replace the appropriate columns of D with the residuals from a regression on all columns to the right of it. Simple inspection shows that the columns of D corresponding to any given effect will already be orthogonal to other effects in D *except* those for interactions that contain it; so the regression does not have to include all columns to the right.

It is easy to demonstrate that this gives the uniform population contrast (Yates) for a large number of data sets, but I have not yet constructed a proof. (I suspect it could be approached using the Rao-Blackwell theorem.)

A.4 Bystanders

What about a model that has a extra predictor, such as `x3` in our example data and in the fit below?

```
> fit4 <- lm(y ~ x1*x2 + x3, data=data1)
```

The standard approach is to ignore this variable when setting up "type III" tests: the contrast for `x1` will be the same as it was in the prior model, with a 0 row in the middle for the x3 coefficient.

A.5 Missing cells

When there are combinations of factors with 0 subjects in that group, it is not possible to create a uniform population via reweighting of either subjects or parameters. There is thus no Yates contrast corresponding to the hypothetical population of interest. For that matter, adjustment to any fixed population is no longer possible, such as the US 2000 reference, unless groups are pooled so as to remove any counts of zero, and even then the estimate could be problematic due to extreme weights.

This fact does not stop each of the above 3 algorithms from executing and producing a number. This raises two further issues. First, what does that number *mean*? Much ink has been spilled on this subject, but I personally have never been able to come to grips with a satisfactory explanation and so have nothing to offer on the topic. I am reluctant to use such estimates. The second issue is that the computational algorithms become more fragile.

- The NSTT algorithm is a disaster in waiting, so no more needs to be said about situations where its behavior may be even worse.
- When fitting the original model, there will be one or more NA coefficients due to the linear dependencies that arise. A natural extension of the ATT method is to leave these out of the sums when computing each average. However, there are data sets for which the particular set of coefficients returned as missing will depend on the order in which variables were listed in the model statement, which in turn will change the ATT result.
- For the STT method, our statement that certain other columns in D will be orthogonal to the chosen effect is no longer true. To match SAS, the orthogonalization step above should include only those effects further to the right that contain the chosen effect (the one we are constructing a contrast vector for). As a side effect, this makes the STT result invariant to the order of the variables in the model statement.

B SAS computations

The following code was executed in version 9.3 of SAS.

```
options ls=70;
libname save "sasdata";

title "Sex only";
proc glm data=save.flc;
    class sex;
    model flc = sex;
title "Sex only";


proc glm data=save.flc;
    class sex age2;
    model flc = age2 sex /solution E1 E2 E3;
title "Second fit, no interaction";


proc glm data=save.flc;
    class sex age2;
    model flc = sex  age2  sex*age2/solution E1 E2 E3;

    estimate 'yates' sex 1 -1  sex*age2 .2 .2 .2 .2 .2 -.2 -.2 -.2 -.2 -.2;

title "Third fit, interaction";

proc phreg data=save.flc;
    class sex age2;
    model futime * death(0) = sex age2/ ties=efron;
```

```
title "Phreg fit, sex and age, additive";

proc phreg data=save.flc;
    class sex age2;
    model futime * death(0) = sex age2  sex*age2 /
          ties=efron type3(all);

    estimate 'Yates sex' sex 1 sex*age2  .2 .2 .2 .2;
    contrast 'NSTT sex ' sex 1 ;
    contrast 'NSTT age' age2 1 0 0 0 ,
                         age2 0 1 0 0 ,
                         age2 0 0 1 0 ,
                         age2 0 0 0 1;
title "Phreg fit, sex and age with interaction";

proc phreg data=save.flc;
    class sex age2/ param=effect;
    model futime * death(0) = sex age2  sex*age2 / ties=efron;
title "Phreg, using effect coding";
```

The SAS output is voluminous, covering over a dozen pages. A subset is extracted below, leaving out portions that are unimportant to our comparison. First the GLM model for sex only. There are no differences between type 1 and type 3 output for this model.

```
...
              Number of Observations Read        7874
              Number of Observations Used        7874
...
Dependent Variable: flc

                                    Sum of
Source                      DF        Squares    Mean Square   F Value

Model                        1      142.19306      142.19306     42.27

Error                     7872    26481.86345        3.36406

Corrected Total           7873    26624.05652
```

The second fit with sex and then age.

```
                  Type I Estimable Functions

                  -----------------Coefficients-----------------
   Effect         age2                                        sex

   Intercept      0                                           0

   age2      1    L2                                          0
   age2      2    L3                                          0
```

age2	3	L4	0
age2	4	L5	0
age2	5	-L2-L3-L4-L5	0
sex	F	-0.2571*L2-0.2576*L3-0.1941*L4-0.0844*L5	L7
sex	M	0.2571*L2+0.2576*L3+0.1941*L4+0.0844*L5	-L7

Type II Estimable Functions

		---Coefficients----	
Effect		age2	sex
Intercept		0	0
age2	1	L2	0
age2	2	L3	0
age2	3	L4	0
age2	4	L5	0
age2	5	-L2-L3-L4-L5	0
sex	F	0	L7
sex	M	0	-L7

Type III Estimable Functions

		---Coefficients----	
Effect		age2	sex
Intercept		0	0
age2	1	L2	0
age2	2	L3	0
age2	3	L4	0
age2	4	L5	0
age2	5	-L2-L3-L4-L5	0
sex	F	0	L7
sex	M	0	-L7

Dependent Variable: flc

Source	DF	Sum of Squares	Mean Square	F Value
Model	5	2212.13649	442.42730	142.60

```
Error                    7868    24411.92003        3.10268

Corrected Total          7873    26624.05652


Source                     DF      Type I SS    Mean Square   F Value

age2                        4    1929.642183     482.410546    155.48
sex                         1     282.494304     282.494304     91.05


Source                     DF     Type II SS    Mean Square   F Value

age2                        4    2069.943424     517.485856    166.79
sex                         1     282.494304     282.494304     91.05


Source                     DF    Type III SS    Mean Square   F Value

age2                        4    2069.943424     517.485856    166.79
sex                         1     282.494304     282.494304     91.05


                                      Standard
Parameter           Estimate             Error    t Value    Pr > |t|

Intercept        5.503757546 B      0.17553667      31.35      <.0001
age2      1     -2.587424744 B      0.17584961     -14.71      <.0001
age2      2     -2.249164537 B      0.17684133     -12.72      <.0001
age2      3     -1.770342603 B      0.17834253      -9.93      <.0001
age2      4     -1.082104827 B      0.18584656      -5.82      <.0001
age2      5      0.000000000 B
sex       F     -0.383454133 B      0.04018624      -9.54      <.0001
sex       M      0.000000000 B
```

The third linear models fit, containing interactions. For first portion I have trimmed off long printout on the right, i.e. the estimable functions for the age2*sex effect since they are not of interest.

```
                    Type I Estimable Functions

                 --------------------Coefficients--------
Effect           sex          age2

Intercept        0            0

sex       F      L2           0
sex       M      -L2          0
```

```
age2      1      -0.0499*L2   L4
age2      2      -0.0373*L2   L5
age2      3      0.0269*L2    L6
age2      4      0.0482*L2    L7
age2      5      0.0121*L2    -L4-L5-L6-L7

sex*age2  F 1    0.3786*L2    0.6271*L4+0.1056*L5+0.0796*L6+0.0346*L7
sex*age2  F 2    0.2791*L2    0.0778*L4+0.5992*L5+0.0587*L6+0.0255*L7
sex*age2  F 3    0.2182*L2    0.0527*L4+0.0528*L5+0.6245*L6+0.0173*L7
sex*age2  F 4    0.1055*L2    0.0188*L4+0.0188*L5+0.0142*L6+0.7006*L7
sex*age2  F 5    0.0186*L2    -0.7764*L4-0.7764*L5-0.777*L6-0.7781*L7
sex*age2  M 1    -0.4285*L2   0.3729*L4-0.1056*L5-0.0796*L6-0.0346*L7
sex*age2  M 2    -0.3164*L2   -0.0778*L4+0.4008*L5-0.0587*L6-0.0255*L7
sex*age2  M 3    -0.1913*L2   -0.0527*L4-0.0528*L5+0.3755*L6-0.0173*L7
sex*age2  M 4    -0.0573*L2   -0.0188*L4-0.0188*L5-0.0142*L6+0.2994*L7
sex*age2  M 5    -0.0065*L2   -0.2236*L4-0.2236*L5-0.223*L6-0.2219*L7

                    Type II Estimable Functions

                  --------------------Coefficients---------------------
Effect            sex          age2

Intercept         0            0

sex       F       L2           0
sex       M       -L2          0

age2      1       0            L4
age2      2       0            L5
age2      3       0            L6
age2      4       0            L7
age2      5       0            -L4-L5-L6-L7

sex*age2  F 1     0.41*L2      0.6271*L4+0.1056*L5+0.0796*L6+0.0346*L7
sex*age2  F 2     0.3025*L2    0.0778*L4+0.5992*L5+0.0587*L6+0.0255*L7
sex*age2  F 3     0.2051*L2    0.0527*L4+0.0528*L5+0.6245*L6+0.0173*L7
sex*age2  F 4     0.073*L2     0.0188*L4+0.0188*L5+0.0142*L6+0.7006*L7
sex*age2  F 5     0.0093*L2    -0.7764*L4-0.7764*L5-0.777*L6-0.7781*L7
sex*age2  M 1     -0.41*L2     0.3729*L4-0.1056*L5-0.0796*L6-0.0346*L7
sex*age2  M 2     -0.3025*L2   -0.0778*L4+0.4008*L5-0.0587*L6-0.0255*L7
sex*age2  M 3     -0.2051*L2   -0.0527*L4-0.0528*L5+0.3755*L6-0.0173*L7
sex*age2  M 4     -0.073*L2    -0.0188*L4-0.0188*L5-0.0142*L6+0.2994*L7
sex*age2  M 5     -0.0093*L2   -0.2236*L4-0.2236*L5-0.223*L6-0.2219*L7


                    Type III Estimable Functions

                 --------------------Coefficients---------------------
Effect           sex      age2                          sex*age2
```

Intercept		0	0		0
sex	F	L2	0		0
sex	M	-L2	0		0
age2	1	0	L4		0
age2	2	0	L5		0
age2	3	0	L6		0
age2	4	0	L7		0
age2	5	0	-L4-L5-L6-L7		0
sex*age2	F 1	0.2*L2	0.5*L4		L9
sex*age2	F 2	0.2*L2	0.5*L5		L10
sex*age2	F 3	0.2*L2	0.5*L6		L11
sex*age2	F 4	0.2*L2	0.5*L7		L12
sex*age2	F 5	0.2*L2	-0.5*L4-0.5*L5-0.5*L6-0.5*L7		-L9-L10-L11-L12
sex*age2	M 1	-0.2*L2	0.5*L4		-L9
sex*age2	M 2	-0.2*L2	0.5*L5		-L10
sex*age2	M 3	-0.2*L2	0.5*L6		-L11
sex*age2	M 4	-0.2*L2	0.5*L7		-L12
sex*age2	M 5	-0.2*L2	-0.5*L4-0.5*L5-0.5*L6-0.5*L7		L9+L10+L11+L12

Source	DF	Type I SS	Mean Square	F Value
sex	1	142.193063	142.193063	45.97
age2	4	2069.943424	517.485856	167.30
sex*age2	4	87.218363	21.804591	7.05

Source	DF	Type II SS	Mean Square	F Value
sex	1	282.494304	282.494304	91.33
age2	4	2069.943424	517.485856	167.30
sex*age2	4	87.218363	21.804591	7.05

Source	DF	Type III SS	Mean Square	F Value
sex	1	126.961986	126.961986	41.05
age2	4	1999.446491	499.861623	161.60
sex*age2	4	87.218363	21.804591	7.05

Parameter	Estimate	Standard Error	t Value	Pr > |t|
yates	-0.58972607	0.09204824	-6.41	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|

Intercept			6.003043478	B	0.36672295	16.37	<.0001
sex	F		-1.024512614	B	0.41553944	-2.47	0.0137
sex	M		0.000000000	B			
age2		1	-3.176876326	B	0.36950532	-8.60	<.0001
age2		2	-2.787597918	B	0.37048599	-7.52	<.0001
age2		3	-2.088127335	B	0.37292760	-5.60	<.0001
age2		4	-1.353746449	B	0.38703805	-3.50	0.0005
age2		5	0.000000000	B			
sex*age2	F	1	0.813889663	B	0.42023749	1.94	0.0528
sex*age2	F	2	0.716160958	B	0.42189464	1.70	0.0896
sex*age2	F	3	0.330651265	B	0.42487846	0.78	0.4365
sex*age2	F	4	0.313230835	B	0.44127621	0.71	0.4778
sex*age2	F	5	0.000000000	B			
sex*age2	M	1	0.000000000	B			
sex*age2	M	2	0.000000000	B			
sex*age2	M	3	0.000000000	B			
sex*age2	M	4	0.000000000	B			
sex*age2	M	5	0.000000000	B			

The phreg printout for the additive model with age and sex.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2357.5239	5	<.0001
Score	3823.3905	5	<.0001
Wald	2374.5250	5	<.0001

Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	69.9646	<.0001
age2	4	2374.5211	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
sex	F	1	-0.36617	0.04378	69.9646	<.0001
age2		1	-4.18209	0.12180	1179.0289	<.0001
age2		2	-3.23859	0.11418	804.5068	<.0001
age2		3	-2.17521	0.10963	393.6524	<.0001
age2		4	-1.15226	0.11072	108.3077	<.0001

The model with age*sex interaction.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	37736.900	35374.050
AIC	37736.900	35392.050
SBC	37736.900	35443.188

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2362.8497	9	<.0001
Score	3873.5113	9	<.0001
Wald	2357.9498	9	<.0001

Type 3 Tests

LR Statistics

Effect	DF	Chi-Square	Pr > ChiSq
sex	1	0.4607	0.4973
age2	4	932.1371	<.0001
sex*age2	4	5.3258	0.2555

Score Statistics

Effect	DF	Chi-Square	Pr > ChiSq
sex	1	0.4757	0.4904
age2	4	1506.8699	<.0001
sex*age2	4	5.2516	0.2624

Wald Statistics

Effect	DF	Chi-Square	Pr > ChiSq
sex	1	0.4833	0.4869
age2	4	964.6007	<.0001
sex*age2	4	5.2322	0.2643

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square

sex	F	1	-0.16537	0.23789	0.4833
age2	1	1	-4.02699	0.22585	317.9171
age2	2	1	-3.04796	0.21843	194.7187
age2	3	1	-1.99577	0.21577	85.5504
age2	4	1	-1.10659	0.22256	24.7216
sex*age2	F 1	1	-0.21121	0.26896	0.6167
sex*age2	F 2	1	-0.29334	0.25518	1.3214
sex*age2	F 3	1	-0.25663	0.24829	1.0684
sex*age2	F 4	1	-0.04339	0.25527	0.0289

Contrast	DF	Chi-Square	Pr > ChiSq
NSTT sex	1	0.4833	0.4869
NSTT age	4	964.6007	<.0001

Likelihood Ratio Statistics for Type 1 Analysis

Source	-2 Log L	DF	LR Chi-Square	Pr > ChiSq
(Without Covariates)	37736.8997			
sex	37733.0932	1	3.8066	0.0511
age2	35379.3758	4	2353.7173	<.0001
sex*age2	35374.0501	4	5.3258	0.2555

Label	Estimate	Standard Error	z Value	Pr > |z|
Yates	-0.3263	0.06149	-5.31	<.0001

References

[1] M. Aitkin (1978). The analysis of unbalanced cross classifications (with discussion). *J Royal Stat Soc A* 141:195-223.

[2] A. Dispenzieri, J. Katzmann, R. Kyle, D. Larson, T. Therneau, C. Colby, R. Clark, .G Mead, S. Kumar, L..J Melton III and S.V. Rajkumar (2012). Use of monoclonal serum immunoglobulin free light chains to predict overall survival in the general population, *Mayo Clinic Proc* 87:512–523.

[3] D. G. Herr (1986). On the History of ANOVA in Unbalanced, Factorial Designs: The First 30 Years. *Amer Statistician* 40:265-270.

[4] R. Kyle, T. Therneau, S.V. Rajkumar, D. Larson, M. Plevak, J. Offord, A. Dispenzieri, J. Katzmann, and L.J. Melton, III (2006), Prevalence of monoclonal gammopathy of undetermined significance, *New England J Medicine* 354:1362–1369.

[5] D. B. Macnaughton (1992). Which sum of squares are best in an unbalanced analysis of variance. www.matstat.com/ss.

[6] J. Nelder (1977). A reformulation of linear models (with discussion). *J Royal Stat Soc A* 140:48–76.

[7] SAS Institute Inc. (2008), The four types of estimable functions. SAS/STAT 9.2 User's Guide, chapter 15.

[8] S. R. Searle, *Linear Models*, Wiley, New York, 1971.

[9] S. Senn. Multi-centre trials and the finally decisive argument. www.senns.demon.co.uk/wprose.html#FDA.

[10] S. Senn. Good mixed centre practice. www.senns.demon.co.uk/wprose.html#Mixed.

[11] S. Senn. Statistical Issues in Drug Development, Wiley, New York, 2007.

[12] S. Senn. The many modes of meta. Drug Information J 34:535-549, 2000.

[13] T. M. Therneau and P. M. Grambsch, *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York, 2000.

[14] F. Yates (1934). The analysis of multiple classifications with unequal numbers in the different classes. *J Am Stat Assoc*, 29:51–66.