# TrainMALTA - Statistics Exercises

Some of the datasets used in this assignment are stored as tables, each of which you can read into R as a dataframe using:

```
> mydata<-read.table("/data/day2/datasets/filename.txt", head=TRUE, stringsAsFactors=FALSE)
```

To make the variables directly available in the workspace, without reference to the dataframe, you can issue the command:

```
> attach(mydata)
```

## Question 1

Consider the univariate logistic regression equation:

$$\text{logit}\, P(Y = 1) = \beta_0 + \beta_1 X$$

and the following four quantities derived from that equation:

$$\frac{e^{\beta_0}}{1 + e^{\beta_0}} \tag{A}$$

$$\frac{e^{\beta_0 + \beta_1 .2}}{1 + e^{\beta_0 + \beta_1 .2}} \tag{B}$$

$$e^{\beta_0} \tag{C}$$

$$\beta_1 \tag{D}$$

$$\beta_0 + \beta_1 .2 \tag{E}$$

Match the letters A to E to the following written descriptions:

□ The probability that Y = 1 when X = 2.

□ The log odds for Y = 1 when X = 2.

□ The probability that Y = 1 when X = 0.

□ The odds in favour of Y = 1 when X = 0.

□ The log odds ratio for Y = 1 comparing a pair of observations with values of X that differ by 1.

B, E, A, C, D

## Question 2

The dataset *modsel.txt* contains a continuous outcome variable `y` and three predictor variables `x1`, `x2` and `x3`.

A two-way contingency table cross-classifying observations of dichotomous variables `u` and `v` is a table of counts of the form

|  | u=0 | u=1 |
|---|---|---|
| v=0 | $a$ | $b$ |
| v=1 | $c$ | $d$ |

Given such data we can estimate the odds ratio for association between `u` and `v` using the formula

$$OR = \frac{ad}{bc}. \qquad (1)$$

(a) Create a new 0/1 variable `yd` by dichotomising the continuous variable `y`, using 0 as a threshold:

```
> yd=as.numeric(y>0)
```

Note that variables `x2` and `x3` are dichotomous. Report three contingency tables cross-classifying each of the possible pairings of the variables `x2`, `x3` and `yd`. e.g. to cross classify `x2` and `x3` you can use the command `table(x2, x3)`.

The R output is as follows:

```
> table(x2,x3)
   x3
x2   0  1
  0 54 43
  1 52 51
> table(yd,x2)
   x2
yd   0  1
  0 73 23
  1 24 80
> table(yd,x3)
   x3
yd   0  1
  0 30 66
  1 76 28
```

(b) Estimate an odds ratio for each of the three contingency tables using the formula given in equation (1).

Estimated odds ratios are respectively: i) 1.2317 ii) 10.5797 iii) 0.1675

(c) Fit univariate logistic regressions of i) `yd` on `x2`, ii) `yd` on `x3`, iii) `x2` on `x3` and report estimates of the odds ratio parameter for each regression.

```
> glm(yd~x2, family="binomial")

Call:  glm(formula = yd ~ x2, family = "binomial")

Coefficients:
(Intercept)           x2
     -1.112        2.359

Degrees of Freedom: 199 Total (i.e. Null);  198 Residual
Null Deviance:     276.9
Residual Deviance: 217.9  AIC: 221.9
> glm(yd~x3, family="binomial")

Call:  glm(formula = yd ~ x3, family = "binomial")

Coefficients:
(Intercept)           x3
     0.9295      -1.7870

Degrees of Freedom: 199 Total (i.e. Null);  198 Residual
Null Deviance:     276.9
Residual Deviance: 240.8  AIC: 244.8
> glm(x2~x3, family="binomial")

Call:  glm(formula = x2 ~ x3, family = "binomial")

Coefficients:
(Intercept)           x3
   -0.03774      0.20837

Degrees of Freedom: 199 Total (i.e. Null);  198 Residual
Null Deviance:     277.1
Residual Deviance: 276.5  AIC: 280.5
```

We can obtain odds-ratio estimates from the log odds-ratio estimates by exponentiating

```
> exp(2.359)
[1] 10.58037
> exp(-1.7870)
[1] 0.1674618
> exp(0.20837)
[1] 1.231669
```

Compare your results to the odds ratio estimates you computed in part (b).

The estimates from the logistic regression models are very similar to the estimates from the two-way contingency tables.

## Question 3

The dataset *drug.txt* contains a table from a study in which a group of researchers are evaluating a new treatment for migraine headaches. A clinical trial was designed to compare new drug A to the standard treatment, drug B. Two hundred subjects were randomised to drug A ($n = 100$) or drug B ($n = 100$) at the onset of their next migraine headache, and each subject reported whether their headache had gone ($Y_i = 1$) three hours later or not ($Y_i = 0$).

Read the table into R and look at the first few lines of the table using:

```
tab = read.table( "/data/day2/datasets/drug.txt", header=TRUE )
head( tab )
attach( tab )
```

(a) Regress the outcome variable y on the treatment variable drug (drug=1 implies treatment with drug A, drug=0 implies treatment with drug B), using a univariate logistic model and report a summary of the fitted model. Include maximum likelihood estimates of the intercept and the regression coefficient and estimates of the associated standard errors.

```
> summary(glm(y~drug, family="binomial"))

Call:
glm(formula = y ~ drug, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5305  -1.2116   0.8615   1.1436   1.1436

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.08004    0.20016   0.400   0.6892
drug         0.72008    0.29464   2.444   0.0145 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 268.37  on 199  degrees of freedom
Residual deviance: 262.29  on 198  degrees of freedom
AIC: 266.29

Number of Fisher Scoring iterations: 4
```

A maximum likelihood estimate of the intercept is 0.08004 with estimated standard error 0.20016. A maximum likelihood estimate of the regression coefficient is 0.72008 with estimated standard error 0.29464

(b) Compute a numerical estimate of the probability that an individual treated with drug A will recover from a migraine within three hours.

The log odds in favour of recovery for someone on drug A can be calculated as $0.08004 + 0.72008 = 0.80012$. We can convert this into a probability using the inverse-logit or expit function $\frac{\exp(0.80012)}{1+\exp(0.80012)} \approx 0.690$ Another way of doing this would be to use the R function predict:

```
> predict(mod, data.frame(drug=1), type="response")
   1
0.69
```

(c) Compute a numerical estimate of the probability that an individual treated with drug B will *fail* to recover from a migraine within three hours.

The log odds in favour of recovery for someone on drug B is 0.08004. We can convert this into a probability of recovery using the inverse-logit or expit function $\frac{\exp(0.08004)}{1+\exp(0.08004)} \approx 0.520$. But we want the probability that the individual fails to recover, which is 1 minus the probability that s/he recovers. i.e. $1 - 0.520 \approx 0.488$. There are other ways this calculation could be done. e.g. we could obtain the log odds in favour of failure to recover by multiplying 0.08004 by $-1$. This can then be converted to a probability via the inverse logit function. Alternatively we could use the R predict function:

```
> 1-predict(mod, data.frame(drug=0), type="response")
   1
0.48
```

(d) Give an interpretation of the regression coefficient corresponding to the drug variable.

Logistic regressions are nonlinear in the probability scale - that is, a specified difference in one of the x variables does not correspond to a constant difference in $P(\ y = 1)$. As a consequence the logistic regression coefficients cannot be directly interpreted on the scale of the data.

The best answer to this question is:

The coefficient is the log-odds ratio (in favour of recovery within three hours) for treatment by drug A compared to treatment by drug B.

However, an interpretation on the odds scale would also do:

$\exp(\beta)$ is the odds-ratio (in favour of recovery within three hours) for treatment by drug A compared to treatment by drug B.

(e) Compute an estimate of a 95% confidence interval for the log-odds ratio for the success of treatment with drug A compared to the success of treatment with drug B.

Using a normal approximation we have:

$0.7208 - 1.96 \times 0.29464 < \beta < 0.7208 + 1.96 \times 0.29464$

so, the interval is

$0.143 < \beta < 1.298$.

(f) Compute an estimate of a 95% confidence interval for the odds ratio for the success of treatment with drug A compared to the success of treatment with drug B.

We exponentiate the limits of the interval above to get an interval on the odds scale:

$\exp(0.143) < \beta < \exp(1.298)$

so the interval is

$1.15 < \beta < 3.67$

(g) Give an estimate of the odds ratio for success of treatment with drug B compared to the success of treatment with drug A.

$\exp(-0.72008) = 0.487$

## Question 4

The dataset *satisfaction.txt* contains a table from a survey that predicts patient satisfaction with hospital services (higher numbers indicate greater satisfaction) given the patient's age at hospital admission, a severity index (higher numbers indicate more severe cases) and an anxiety index (higher numbers indicate more anxiety). Read the table into R and look at the first few lines of the table using:

```
sat = read.table( "/data/day2/datasets/satisfaction.txt", header=TRUE )
head( sat )
```

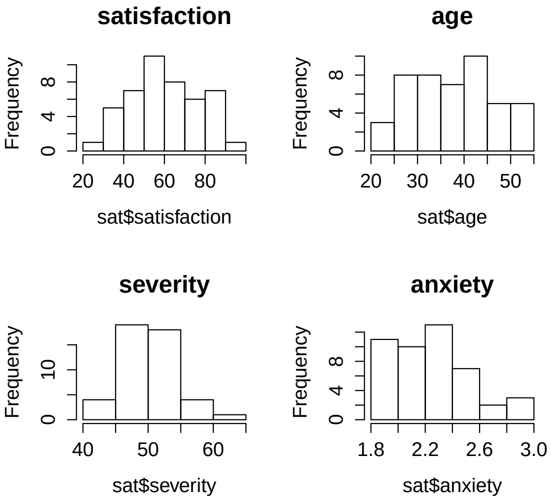
```
  satisfaction age severity anxiety
1           48  50       51     2.3
2           57  36       46     2.3
3           66  40       48     2.2
4           70  41       44     1.8
5           89  28       43     1.8
6           36  49       54     2.9
```

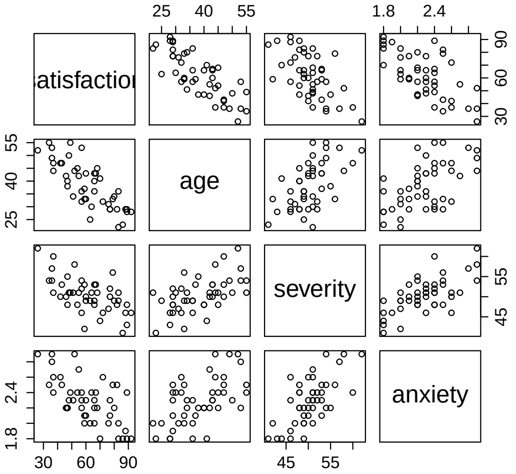
(a) Create histograms of all four variables (use the function `hist`). Note the general features of each variable.

```
hist( sat$satisfaction, main="satisfaction" )
hist( sat$age, main="age" )
hist( sat$severity, main="severity" )
hist( sat$anxiety, main="anxiety" )
```

(b) Use the `pairs` function to look at scatter plots of all possible pairs of variables. Hint: try passing the entire table as an argument to the function. Summarise your findings.

```
pairs( sat )
```


satisfaction
Frequency
0
4
8
20
40
60
80
sat$satisfaction
age
Frequency
0
4
8
20
30
40
50
sat$age
severity
Frequency
0
10
40
50
60
sat$severity
anxiety
Frequency
0
4
8
1.8
2.2
2.6
3.0
sat$anxiety


satisfaction
age
severity
anxiety
25
40
55
1.8
2.4
90
60
30
55
40
25
55
45
2.4
1.8
30
60
90
45
55

There seems to be a relationship between all variables. Satisfaction is negatively correlated with age, severity and anxiety. Age, severity and anxiety appear positively correlated with one another.

(c) Fit a linear regression for each variable separately. Report all parameter estimates.

```
fit.1 = lm( sat$satisfaction ~ sat$age )
summary(fit.1)
fit.2 = lm( sat$satisfaction ~ sat$severity )
summary(fit.1)
fit.3 = lm( sat$satisfaction ~ sat$anxiety )
summary(fit.1)
```

The regression lines are:
satisfaction = 119.9 - 1.5 age + error
satisfaction = 183.1 - 2.4 severity + error
satisfaction = 146.4 - 37 anxiety + error

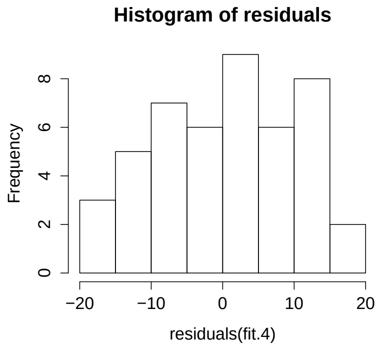
(d) Fit a multiple linear regression for all three variables. Report all parameter estimates.

```
fit.4 = lm( sat$satisfaction ~ sat$age + sat$severity + sat$anxiety )
summary(fit.4)
```

satisfaction = 158.5 - 1.1 age - 0.4 severity - 13.5 anxiety

(e) Plot a histogram of the residuals from the model in (d, use the function `residuals` to extract the values from the fit). Does it look like any assumptions are being violated?

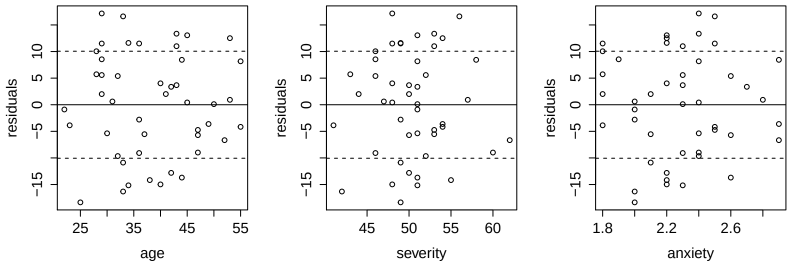
```
hist(residuals(fit.4))
```



The histogram is vaguely normal in shape, but then again, there are just 46 observations. Normality may or may not be met, but this is probably not fatal (the model is robust to non-normality to a certain extent anyway, especially if it remains symmetric).

(f) Create a scatter plot of the residuals agains each of the predictors (so three plots). Comment on what these plots indicate.

```
par(mfrow=c(1,3))
plot( sat$age, residuals(fit.4), xlab="age", ylab="residuals" )
abline(h=0)
abline(h=summary(fit.4)$sigma,lty=2)
abline(h=-summary(fit.4)$sigma,lty=2)
plot( sat$severity, residuals(fit.4), xlab="severity", ylab="residuals" )
abline(h=0)
abline(h=summary(fit.4)$sigma,lty=2)
abline(h=-summary(fit.4)$sigma,lty=2)
plot( sat$anxiety, residuals(fit.4), xlab="anxiety", ylab="residuals" )
abline(h=0)
abline(h=summary(fit.4)$sigma,lty=2)
abline(h=-summary(fit.4)$sigma,lty=2)
```



The scatter plots look reasonable, i.e., no obvious departures from constant variance or linearity. So we have no strong reason to investigate other models. The above conclusions are hand-waving, there exist more formal tests of model fit.

## Question 5

Consider the dataset *plasma.txt*. The data consist of plasma levels of a polyamine (plasma variable $Y$), against age in children ($X$ variable, age $= 0$ indicates a newborn). Read the table into R and look at the first few lines of the table using:

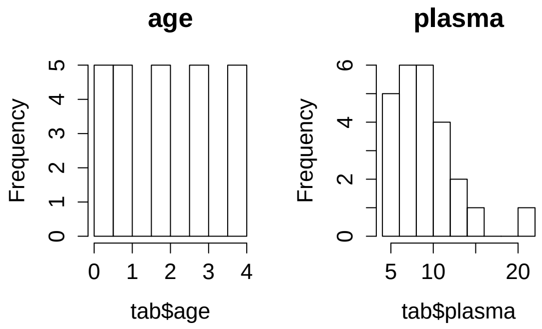
```
tab = read.table( "/data/day2/datasets/plasma.txt", header=TRUE )
head( tab )
```

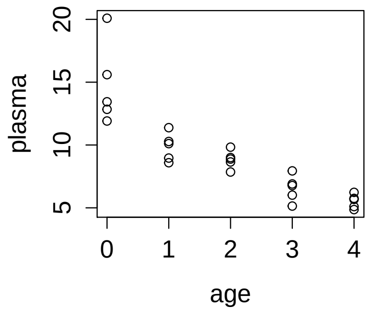
(a) Create histograms for both variables. Note the general features of each variable.

```
hist( tab$age )
hist( tab$plasma )
```

(b) Create a scatter plot of age versus plasma. Does the relationship seem linear?

```
plot( tab$age, tab$plasma, xlab="age", ylab="plasma" )
```





No, it looks like one quantity varies as a power of the other. Perhaps the relationship would be better modelled in the log scale.

(c) Transform the $Y$ variable with the log transform (use the function `log`). In other words, rather than $Y$, create a $log(Y) = log(plasma)$. The logarithm should be to the base $e$. Re-create the scatter plot, but now plotting age versus log(plasma). Does the relationship seem more linear now?

```
log.plasma = log(tab$plasma)
plot( tab$age, log.plasma, xlab="age", ylab="log plasma" )
```

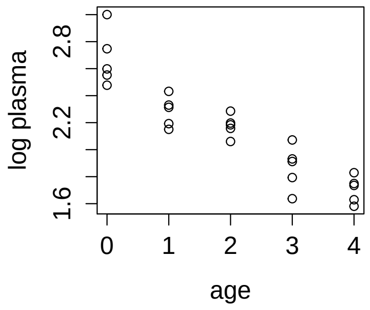
(d) Fit a linear regression for age versus log(plasma). Report all parameter estimates.

```
fit.5 = lm( log.plasma ~ tab$age )
summary(fit.5)
```

plasma = 2.6 - 0.2 age + error

(e) Provide an interpretation of the coefficients calculated in (d).

The estimated coefficient of -0.2 for age implies that a difference in 1 year of age corresponds to an expected negative difference of 0.2 in log(plasma).



## Question 6

The dataset *nonmelanoma.txt* contains counts of the diagnosed non-melanoma skin cancers amongst women in the cities of Dallas-Fort Worth and Minneapolis-St Paul over a period of one year. The counts are grouped by age. The pop variable gives the number of women in each age category in each of the two cities.

(a) Explain why the dataset would be useless as a source of information about the relative rates of non-melanoma skin cancer in Dallas-Fort Worth and Minneapolis-St Paul if the pop variable was missing and assuming we knew nothing about the demographics of the two cities from any other source.

The rates we want to compare measure the expected number of cases of non-melanoma skin cancer in units of person-time:

$$\text{rate} = \frac{\text{expected number of cases}}{\text{amount of person time}}$$

In order to estimate and compare such rates we need to know the size of each city, otherwise we do not have enough information to compute the denominator of the above fraction.

(b) Fit a Poisson regression, regressing cases on agegrp and city, treating both the predictors as R factor variables.

```
> pois1.obj<-glm(cases~offset(log(pop))+agegrp+city,family="poisson")
> summary(pois1.obj)

Call:
glm(formula = cases ~ offset(log(pop)) + agegrp + city, family = "poisson")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.49752  -0.46511   0.01901   0.37143   1.26185

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -10.85314    0.44748 -24.254  < 2e-16 ***
agegrpage25-34   2.63015    0.46746   5.626 1.84e-08 ***
agegrpage35-44   3.84730    0.45466   8.462  < 2e-16 ***
```

```
agegrpage45-54   4.59514    0.45103  10.188  < 2e-16 ***
agegrpage55-64   5.08725    0.45030  11.297  < 2e-16 ***
agegrpage65-74   5.64542    0.44975  12.552  < 2e-16 ***
agegrpage75-84   6.05860    0.45032  13.454  < 2e-16 ***
agegrpage85+     6.19435    0.45774  13.532  < 2e-16 ***
cityminnes      -0.80581    0.05221 -15.435  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2793.0241  on 15  degrees of freedom
Residual deviance:    7.9532  on  7  degrees of freedom
AIC: 120.2

Number of Fisher Scoring iterations: 4

>
```

(c) Referring to the R summary output for the model object comment on the relationship between non-melanoma skin cancer rates and age, within the cities.

The estimated rate ratios comparing each age category with the reference category (age 15-24) increase systematically with age. i.e. there is a trend.

Additional comment: The trend must be significant at the $95\%$ level since the rate in age category 25-34 is significantly higher than the rate in the reference category. The estimates for the higher age categories increase monotonically so only add additional evidence in favour of the hypothesis that there is a trend.

(d) Give estimates of the rates of non-melanoma skin cancer cases amongst 15-24 year old women in both Dallas-Fort Worth and Minneapolis-St Paul, together with estimates of corresponding 95% confidence intervals.

In Dallas-Fort Worth:

$\exp(-10.85314)$ per woman year $= 1.934377 \times 10^{-05}$ per woman year

Lower limit of confidence interval:

$\exp(-10.85314-1.96\times 0.44748)$ per woman year $= 8.047084\times 10^{-06}$ per woman year.

Upper limit of confidence interval:

$\exp(-10.85314+1.96\times 0.44748)$ per woman year $= 4.649903\times 10^{-05}$ per woman year.

For Minne-St Paul, fit a second model:

```
> pois2.obj<-glm(cases~offset(log(pop))+agegrp+city.dallas,family="poisson")
> summary(pois2.obj)

Call:
glm(formula = cases ~ offset(log(pop)) + agegrp + city.dallas,
    family = "poisson")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.49752  -0.46511   0.01901   0.37143   1.26185

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -11.65895    0.44871 -25.983  < 2e-16 ***
```

```
agegrpage25-34   2.63015    0.46746   5.626 1.84e-08 ***
agegrpage35-44   3.84730    0.45466   8.462  < 2e-16 ***
agegrpage45-54   4.59514    0.45103  10.188  < 2e-16 ***
agegrpage55-64   5.08725    0.45030  11.297  < 2e-16 ***
agegrpage65-74   5.64542    0.44975  12.552  < 2e-16 ***
agegrpage75-84   6.05860    0.45032  13.454  < 2e-16 ***
agegrpage85+     6.19435    0.45774  13.532  < 2e-16 ***
city.dallas      0.80581    0.05221  15.435  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2793.0241  on 15  degrees of freedom
Residual deviance:    7.9532  on  7  degrees of freedom
AIC: 120.2

Number of Fisher Scoring iterations: 4
```

$\exp(-11.65895)$ per woman year $= 8.641365 \times 10^{-06}$ per woman year

Lower limit of confidence interval:

$\exp(-11.65895{-}1.96{\times}0.44748)$ per woman year $= 3.586185{\times}10^{-06}$ per woman year.

Upper limit of confidence interval:

$\exp(-11.65895{+}1.96{\times}0.44748)$ per woman year $= 2.082246{\times}10^{-05}$ per woman year.

It would be clearer to present all the above numbers by multiplying them by a large number e.g. 100 000 and adjusting the units to reflect this change.

(e) What are the units of the rates you estimated in part (d)?

Any choice of units from: ”per person-year” ”(person-year)$^{-1}$” ”cases per person-year” is acceptable. To be very precise the units are per woman-year (where a woman is a person over 15 years). If you multiplied the rates estimated in (d) by a large number e.g. (100 000) to make them more human friendly the units will be different (e.g per 100 000 person years).

(f) Give an estimate of the rate ratio comparing the rate of non-melanoma skin cancer in Dallas-Fort Worth with that in Minneapolis-St Paul, within groups of women of the same age. Give an estimate of the corresponding 95% confidence interval.

The maximum likelihood estimate of the ratio:

$\exp(0.80581)$

The lower limit of the confidence interval:

$\exp(0.80581 - 1.96 * 0.05221) = 2.02077$

The upper limit of the confidence interval:

$\exp(0.80581 + 1.96 * 0.05221) = 2.47971$

(g) What are the units of the rate ratio you estimated in part (f)?

Rate ratios are unitless.