Discrete analysis_HW3

ID : 111024517 Name : 鄭家豪

Problem 1

The data come from a study of breast cancer in Wisconsin (Bennet and Mangasarian, 1992). There are 681 cases of potentially cancerous tumors of which 238 are actually malignant. Determining whether a tumor is really malignant is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure, called *fine needle aspiration*, which draws only a small sample of tissue could be effective in determining tumor status. The data has the following response and nine predictors:

- Class - 0 if malignant, 1 if benign
- Adhes - marginal adhesion
- BNucl - bare nuclei
- Chrom - bland chromatin
- Epith - epithelial cell size
- Mitos - mitoses
- Nnucl - normal nucleoli
- Thick - clump thickness
- UShap - cell shape uniformity
- USize - cell size uniformity

The predictor values are determined by a doctor observing the cells and rating them on a scale from 1 (normal) to 10 (most abnormal) with respect to the particular characteristic.

a. Fit a binomial regression with Class as the response and the other nine variables as predictors. Report the residual deviance and associated degrees of freedom. Can this information be used to determine if this model fits the data? Explain.
b. Use AIC as criterion to determine the best subset of variables. (Use the step() function).
c. Use the reduced model to predict the outcome for a new patient with predictor variables 1,1,3,2,1,1,4,1,1 (same order as above). Give a confidence interval for your prediction.
d. Suppose that a cancer is classified as benign if p>0.5 and malignant if p<0.5. Compute the number of errors of both types that will be made if this method is applied to the current data with the reduced model.
e. Suppose we change the cutoff to 0.9 so that p<0.9 is classified as malignant and p>0.9 as benign. Computer the number of errors in this case. Discuss the issues in determining the cutoff.
f. It is usually misleading to use the same data to fit a model and test its predictive ability. To investigate this, split the data into two parts - assign every third observation to a test set and the remaining 2/3 of the data to a training set. Use the training set to determine the model and the test set to assess its predictive performance. Compare the outcome to the previously obtained results.

Sol.

(a)

Use data to fit the following model:
$\text{logit}(p_x) \sim 1+$ Adhes$+$ BNucl$+$ Chrom$+$ Epith$+$ Mitos$+$ Nuncl$+$ Thick$+$ UShap$+$ USize , where $p_x = class_x/n_x$.

```
data = read.table("wbca.txt",header = T)
dim(data)
```

```
[1] 681  10
```

```
fit1 = glm(Class ~ . , data = data, family ="binomial")
summary(fit1)
```

```
Call:
glm(formula = Class ~ ., family = "binomial", data = data)

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.16678    1.41491   7.892 2.97e-15 ***
Adhes       -0.39681    0.13384  -2.965  0.00303 **
BNucl       -0.41478    0.10230  -4.055 5.02e-05 ***
Chrom       -0.56456    0.18728  -3.014  0.00257 **
Epith       -0.06440    0.16595  -0.388  0.69795
Mitos       -0.65713    0.36764  -1.787  0.07387 .
NNucl       -0.28659    0.12620  -2.271  0.02315 *
Thick       -0.62675    0.15890  -3.944 8.01e-05 ***
UShap       -0.28011    0.25235  -1.110  0.26699
USize        0.05718    0.23271   0.246  0.80589
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 881.388  on 680  degrees of freedom
Residual deviance:  89.464  on 671  degrees of freedom
AIC: 109.46

Number of Fisher Scoring iterations: 8
```

Based on this result, we can conclude that this model adequately fits the data as evidenced by the small residual deviance. However, due to the high degree of freedom under the null hypothesis (680), indicating sparsity in the data, using deviance as a measure of goodness-of-fit may not be suitable.

(b)

```
fit2 = step(fit1, trace = 0)
summary(fit2)
```

```
Call:
glm(formula = Class ~ Adhes + BNucl + Chrom + Mitos + NNucl +
    Thick + UShap, family = "binomial", data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  11.0333     1.3632   8.094 5.79e-16 ***
Adhes        -0.3984     0.1294  -3.080  0.00207 **
BNucl        -0.4192     0.1020  -4.111 3.93e-05 ***
Chrom        -0.5679     0.1840  -3.085  0.00203 **
Mitos        -0.6456     0.3634  -1.777  0.07561 .
NNucl        -0.2915     0.1236  -2.358  0.01837 *
Thick        -0.6216     0.1579  -3.937 8.27e-05 ***
UShap        -0.2541     0.1785  -1.423  0.15461
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 881.388  on 680  degrees of freedom
```

```
Residual deviance:  89.662  on 673  degrees of freedom
AIC: 105.66

Number of Fisher Scoring iterations: 8
```

By AIC as criterion, the best subset of models is
$\text{logit}(p_x) \sim 1+$ Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap , where $p_x = class_x/n_x$.

(c)

```
pred.p = matrix( c(1, 1, 3, 2, 1, 1, 4, 1, 1),nrow = 1)
pred.p = as.data.frame(pred.p)
colnames(pred.p) <- colnames(data)[-1]
pred = predict(fit2, newdata=pred.p,se=T)
```

The predicted value for a new patient with (1,1,3,2,1,1,4,1,1) is

```
as.numeric(pred$fit)
```

```
[1] 4.834428
```

And, the 95 % confidence interval for this prediction is

```
CI.pred = pred$fit + c(-1,1) * qnorm(0.975) * pred$se.fit
CI.pred
```

```
[1] 3.694673 5.974183
```

Moreover, its predicted benign probability and 95 % confidence interval for benign probability are, respectively,

```
library(faraway)
as.numeric( ilogit( pred$fit) )
```

```
[1] 0.9921115
```

```
as.numeric( ilogit( CI.pred ) )
```

```
[1] 0.9757472 0.9974629
```

(d)

```
est.prob = predict(fit2, newdata = data, type = "response")
est.class = as.factor( ifelse( est.prob > 0.5, 1, 0) )
result = table(est.class, data$Class)
result
```

```
est.class   0   1
        0 227   9
        1  11 434
```

```
TypeI = as.numeric( result[1,2]/(colSums(result)[2]) )
TypeII = as.numeric( result[2,1]/(colSums(result)[1]) )
cat("P(Type I)=", TypeI, "\n")
```

```
P(Type I)= 0.02031603
```

```
cat("P(Type II)=", TypeII)
```

```
P(Type II)= 0.04621849
```

(e)

```
est.class2 = as.factor( ifelse( est.prob > 0.9, 1, 0) )
result2 = table(est.class2, data$Class)
result2
```

```
est.class2   0   1
         0 237  16
         1   1 427
```

```
TypeI2 = as.numeric( result2[1,2]/(colSums(result2)[2]) )
TypeII2 = as.numeric( result2[2,1]/(colSums(result2)[1]) )
cat("P(Type I)=", TypeI2, "\n")
```

```
P(Type I)= 0.03611738
```

```
cat("P(Type II)=", TypeII2)
```

```
P(Type II)= 0.004201681
```

When changing the cutoff from 0.5 to 0.9, the probability of type I error will be increased, while the probability of type II error will be decreased. In the meaning of breast cancer screeing, reducing the probability of the type II error is crucial. Therefore, a cutoff value of 0.9 is preferable.

(f)

Based on the conclusion of part (e), the cutoff value is setting 0.9, and we proceed with the train-test analysis:

```
test.index = seq(3, 681, 3)
train.data = data[-test.index, ]
test.data = data[test.index, ]

fit_train = glm(Class~., data= train.data, family = "binomial")
step.fit_train = step(fit_train, trace = 0)
summary(step.fit_train)
```

```
Call:
glm(formula = Class ~ Adhes + BNucl + Chrom + Mitos + NNucl +
    Thick + UShap, family = "binomial", data = train.data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  11.5571     1.8285   6.321  2.6e-10 ***
Adhes        -0.4249     0.1441  -2.949  0.00318 **
BNucl        -0.3341     0.1187  -2.815  0.00487 **
Chrom        -0.5963     0.2422  -2.462  0.01382 *
Mitos        -0.5822     0.4872  -1.195  0.23207
NNucl        -0.4192     0.1604  -2.614  0.00895 **
Thick        -0.6037     0.1924  -3.138  0.00170 **
UShap        -0.2943     0.2034  -1.447  0.14795
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 592.796  on 453  degrees of freedom
Residual deviance:  59.536  on 446  degrees of freedom
AIC: 75.536

Number of Fisher Scoring iterations: 9
```

The fitted model of train dataset: $\text{logit}(p_x) \sim 1+ \text{Adhes} + \text{BNucl} + \text{Chrom} + \text{Mitos} + \text{NNucl} + \text{Thick} + \text{UShap}$, where $p_x = class_x/n_x$.

Summarized the type I and type II error of train dataset:

```
est.prob_train = predict(step.fit_train, newdata = train.data, type = "response")
est.class_train = as.factor( ifelse( est.prob_train > 0.9, 1, 0) )
result_train = table(est.class_train, train.data$Class)
result_train
```

```
est.class_train   0   1
              0 163  14
              1   0 277
```

```
TypeI.train = as.numeric( result_train[1,2]/(colSums(result_train)[2]) )
TypeII.train = as.numeric( result_train[2,1]/(colSums(result_train)[1]) )
cat("P(Type I in train)=", TypeI.train, "\n")
```

```
P(Type I in train)= 0.04810997
```

```
cat("P(Type II in train)=", TypeII.train)
```

```
P(Type II in train)= 0
```

Next, we use this model to predict testing data:

```
pred.prob_test = predict(step.fit_train, newdata = test.data, type = "response")
pred.class_test = as.factor( ifelse( pred.prob_test > 0.9, 1, 0) )
result_test = table(pred.class_test, test.data$Class)
result_test
```

```
pred.class_test   0   1
              0  73   3
              1   2 149
```

```
TypeI.test = as.numeric( result_test[1,2]/(colSums(result_test)[2]) )
TypeII.test = as.numeric( result_test[2,1]/(colSums(result_test)[1]) )
cat("P(Type I in test)=", TypeI.test, "\n")
```

```
P(Type I in test)= 0.01973684
```

```
cat("P(Type II in test)=", TypeII.test)
```

```
P(Type II in test)= 0.02666667
```

Comparing the type I and type II error probability between the training and testing sets, we observe that the type I error probability in the test set is lower than in the training set, while the type II error probability increases from 0% in the training set to a non-zero value in the test set. This suggests that as the data is divided into training and testing sets, the phenomenon of information dispersion reflects in Error rate.

Problem 2

The dataset infert presents data from a study of secondary infertility (failure to conceive after at least one previous conception). The variables in the dataset are:

- education - 0=0-5 years; 1=6-11 years; 2=12+ years
- age - age in years of case
- parity - number of prior pregnancies
- induced - number of prior induced abortions, 0=0; 1=1; 2=2 or more
- case - case status, 1=case; 0=control
- spontaneous - number of prior spontaneous abortions, 0=0; 1=1; 2=2 or more
- stratum - matched set number, 1-83
- pooled.stratum - stratum number, 1-63

The factors of interest are induced abortions and spontaneous abortions (e.g., miscarriages). The study matched each case of infertility with two controls who were not infertile, matching on age, education, and parity.

a. Construct cross-classified tables by number of spontaneous and induced abortions separately for cases and controls. Comment on the differences between the two tables.
b. Fit a binary response model with only spontaneous and induced as predictors. Determine the statistical significance of these predictors. Express the effects of the predictors in terms of odds.
c. Fit a binary response model with only education, age, and parity as predictors. Explain how the significance (or lack thereof) of these predictors should be interpreted.
d. Now put all five predictors in a binary response model. Interpret the results in terms of odds.
e. Fit a matched case control model appropriate to the data. Interpret the output and compare the odds to those found in the previous model.
f. The spontaneous and induced predictors could be viewed as ordinal due to the grouping in the highest level. Refit the model using ordinal factors rather than numerical variables for these two predictors. Is there evidence that the ordinal representation is necessary?

Sol.

```
data = infert
data$education = factor(data$education, labels = c(0,1,2))
str(data)
```

```
'data.frame':   248 obs. of  8 variables:
 $ education     : Factor w/ 3 levels "0","1","2": 1 1 1 1 2 2 2 2 2 2 ...
 $ age           : num  26 42 39 34 35 36 23 32 21 28 ...
 $ parity        : num  6 1 6 4 3 4 1 2 1 2 ...
 $ induced       : num  1 1 2 2 1 2 0 0 0 0 ...
```

```
 $ case          : num  1 1 1 1 1 1 1 1 1 1 ...
 $ spontaneous   : num  2 0 0 0 1 1 0 0 1 0 ...
 $ stratum       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ pooled.stratum: num  3 1 4 2 32 36 6 22 5 19 ...
```

(a)

```
case.index = data$case == 1
control.index = data$case == 0
cases_table <- table("spontaneous"=data$spontaneous[case.index],
                             "induced"=data$induced[case.index])
cat("Cross-classified table for cases:","\n") ; cases_table
```

```
Cross-classified table for cases:

           induced
spontaneous  0  1  2
          0  7 12  9
          1 22  5  4
          2 18  6  0
```

```
controls_table <- table("spontaneous"=data$spontaneous[-case.index],
                             "induced"=data$induced[-case.index])
cat("Cross-classified table for controls:","\n") ; controls_table
```

```
Cross-classified table for controls:

           induced
spontaneous  0  1  2
          0 67 45 29
          1 47 16  8
          2 29  6  0
```

Now, to comment on the differences between the two tables, looking for any notable patterns or discrepancies that stand out, such as lower frequencies in certain categories for cases compared to controls or vice versa. These differences could provide insights into potential associations or risk factors related to spontaneous and induced abortions. If we want to fit the statistical model, these variables may be useful.

(b)

```
fit = glm(case ~ spontaneous + induced,data = data, family = "binomial")
summary(fit)
```

```
Call:
glm(formula = case ~ spontaneous + induced, family = "binomial",
    data = data)

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7079     0.2677  -6.380 1.78e-10 ***
spontaneous   1.1972     0.2116   5.657 1.54e-08 ***
induced       0.4181     0.2056   2.033    0.042 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 316.17  on 247  degrees of freedom
Residual deviance: 279.61  on 245  degrees of freedom
AIC: 285.61

Number of Fisher Scoring iterations: 4
```

From the above, with the significant level 0.05, only spontaneous is significant. We may conclude that this model is adequate because the residual deviance 279.61 is not larger than $\chi^2_{0.95,245} = 282.5115$. Next, start with the effect of two predictors in odds scale:

```
cat("spontaneous effect:",exp(1.1972), "\n") ; cat("induced effect:",exp(0.4181))
```

```
spontaneous effect: 3.310834

induced effect: 1.519073
```

For each unit increaseing in spontaneous (up to 2 units), the odds of infertility increase approximately 3.31 times. And for each unit increasing in induced (up to 2 units), the odds of infertility increase approximately 1.52 times.

(c)

```
fit2 = glm(case ~ education + age + parity,data = data, family = "binomial")
summary(fit2)
```

```
Call:
glm(formula = case ~ education + age + parity, family = "binomial",
    data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.847542   1.245895  -0.680    0.496
education1   0.046079   0.693754   0.066    0.947
education2   0.069988   0.718125   0.097    0.922
age          0.002076   0.027245   0.076    0.939
parity       0.019070   0.117221   0.163    0.871

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 316.17  on 247  degrees of freedom
Residual deviance: 316.14  on 243  degrees of freedom
```

```
AIC: 326.14

Number of Fisher Scoring iterations: 4
```

These predictors are all insignificant at level 0.05, and the residual deviance 316.14 is larger than $\chi^2_{0.95,243} = 280.3624$, this model is inadequate to fit this data.

(d)

```
fit3 = glm(case ~ education + age + parity + spontaneous + induced,
           data = data, family = "binomial")
summary(fit3)
```

```
Call:
glm(formula = case ~ education + age + parity + spontaneous +
    induced, family = "binomial", data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.14924    1.41220  -0.814   0.4158
education1  -1.04424    0.79255  -1.318   0.1876
education2  -1.40321    0.83416  -1.682   0.0925 .
age          0.03958    0.03120   1.269   0.2046
parity      -0.82828    0.19649  -4.215 2.49e-05 ***
spontaneous  2.04591    0.31016   6.596 4.21e-11 ***
induced      1.28876    0.30146   4.275 1.91e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 316.17  on 247  degrees of freedom
Residual deviance: 257.80  on 241  degrees of freedom
AIC: 271.8

Number of Fisher Scoring iterations: 4
```

Fit the model: logit(case) ~ education + age + parity + spontaneous + induced, with the significant level 0.05, the significant variables are parity, spontaneous, induced. And the residual deviance 257.80 is not larger than $\chi^2_{0.95,241} = 278.2127$, we do not reject this model.

In terms of odds,

```
cat("parity effect:",exp(-0.82828 ), "\n") ;
```

```
parity effect: 0.4367999
```

```
cat("spontaneous effect:",exp(2.04591 ), "\n") ;
```

```
spontaneous effect: 7.736195
```

```
cat("induced effect:",exp(1.28876 ))
```

```
induced effect: 3.628285
```

which means:

- For unit increaseing in parity(up to 6 units), the odds of infertility reduces approximately 43.68%.
- For unit increaseing in spontaneous(up to 2 units), the odds of infertility increases approximately 7.74 times.
- For unit increaseing in induced(up to 2 units), the odds of infertility increases approximately 3.63 times.

In part (b), "induced" variable was not significant. However, the results above indicate that "induced" is a significant variable, with larger effects on odds compared to the model from part (b). This suggests that "parity" are linked to some unknown confounding variables.

(e)

First, we should check if there are any missing items in the matching set of this data.

```
as.numeric( which( table(data$stratum) != 3 ) )
```

```
[1] 74
```

For the Matched case-control design analysis, we should delete the 74-th observation, and rename the data deleted it as data_m.

```
dele = which(data$stratum == 74)
data_m = data[- dele, ]
library(survival)
cmod <- clogit(case ~ education+age+parity+induced+spontaneous, data_m)
summary(cmod)
```

```
Call:
coxph(formula = Surv(rep(1, 246L), case) ~ education + age +
    parity + induced + spontaneous, data = data_m, method = "exact")

  n= 246, number of events= 82

               coef exp(coef) se(coef)      z Pr(>|z|)
education1  -1.20803   0.29878  0.81058 -1.490   0.1361
education2  -1.66714   0.18879  0.86191 -1.934   0.0531 .
age          0.03392   1.03450  0.03139  1.081   0.2799
parity      -0.93991   0.39066  0.21353 -4.402 1.07e-05 ***
induced      1.40290   4.06697  0.31518  4.451 8.54e-06 ***
spontaneous  2.10662   8.22041  0.31864  6.611 3.81e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
education1     0.2988     3.3469   0.06101    1.4633
education2     0.1888     5.2970   0.03486    1.0224
age            1.0345     0.9666   0.97277    1.1001
parity         0.3907     2.5597   0.25707    0.5937
induced        4.0670     0.2459   2.19275    7.5431
```

```
spontaneous    8.2204     0.1216   4.40220   15.3503

Concordance= 0.787  (se = 0.032 )
Likelihood ratio test= 59.92  on 6 df,   p=5e-11
Wald test            = 43.77  on 6 df,   p=8e-08
Score (logrank) test = 54.77  on 6 df,   p=5e-10
```

From the above result,

• The odds of infertility increase by a factor of 4.067 for each unit increase in the induced (up to 2), which is not far away from the 3.628285 we got in the result of part(d), but is still somewhat different. Other significant variables, parity and spontaneous, have same comments.

Let us recursively eliminate predictors/effects which are not significant by using step() in R:

```
cmod1 = step(cmod,trace=0)
summary(cmod1)
```

```
Call:
coxph(formula = Surv(rep(1, 246L), case) ~ education + parity +
    induced + spontaneous, data = data_m, method = "exact")

  n= 246, number of events= 82

               coef exp(coef) se(coef)      z Pr(>|z|)
education1  -1.3078    0.2704   0.8020 -1.631   0.1030
education2  -1.8618    0.1554   0.8409 -2.214   0.0268 *
parity      -0.9269    0.3958   0.2126 -4.359 1.31e-05 ***
induced      1.3633    3.9091   0.3126  4.361 1.29e-05 ***
spontaneous  2.0634    7.8725   0.3146  6.559 5.40e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
education1     0.2704     3.6979   0.05615    1.3023
education2     0.1554     6.4355   0.02990    0.8076
parity         0.3958     2.5268   0.26088    0.6004
induced        3.9091     0.2558   2.11838    7.2135
spontaneous    7.8725     0.1270   4.24966   14.5839

Concordance= 0.785  (se = 0.032 )
Likelihood ratio test= 58.74  on 5 df,   p=2e-11
Wald test            = 43.09  on 5 df,   p=4e-08
Score (logrank) test = 53.67  on 5 df,   p=2e-10
```

The conclusion of selected variables from Surv(rep(1, 246L), case) ~ education + parity + induced + spontaneous is the same with those from Surv(rep(1, 246L), case) ~ education + age + parity + induced + spontaneous at level 0.01. If we set level 0.05, the education is marginal significant, here we omit it.

(f)

```
cmod.order <- clogit(case ~ as.ordered(induced) + as.ordered(spontaneous), data_m)
summary(cmod.order)
```

```
Call:
coxph(formula = Surv(rep(1, 246L), case) ~ as.ordered(induced) +
    as.ordered(spontaneous), data = data_m, method = "exact")

  n= 246, number of events= 82

                             coef exp(coef) se(coef)      z Pr(>|z|)
as.ordered(induced).L      0.57903   1.78430  0.30483  1.900   0.0575 .
as.ordered(induced).Q     -0.03513   0.96548  0.28117 -0.125   0.9006
as.ordered(spontaneous).L  1.68325   5.38302  0.31712  5.308 1.11e-07 ***
as.ordered(spontaneous).Q -0.07532   0.92745  0.26252 -0.287   0.7742
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                          exp(coef) exp(-coef) lower .95 upper .95
as.ordered(induced).L        1.7843     0.5604    0.9817     3.243
as.ordered(induced).Q        0.9655     1.0358    0.5564     1.675
as.ordered(spontaneous).L    5.3830     0.1858    2.8913    10.022
as.ordered(spontaneous).Q    0.9275     1.0782    0.5544     1.551

Concordance= 0.729  (se = 0.033 )
Likelihood ratio test= 36.5  on 4 df,   p=2e-07
Wald test            = 31.95  on 4 df,   p=2e-06
Score (logrank) test = 36.29  on 4 df,   p=3e-07
```

Based on the fitting results converting induced and spontaneous as to ordinal types, we observe that the linear effect is significantly more pronounced than the quadratic effect for these two variables. Therefore, we conclude that an ordinal representation is unnecessary.