Exercises from: Models for Matched Pairs

Chapter 8 Agresti (2007)
Saúl Sotomayor Leytón
April 2012

Setup

```
> library(vcd)
> library(survival)
> library(exact2x2)
> library(BradleyTerry2)
> options(width=70)
```

Problem 1

Remember that the `mcnemar.test` needs the input as a matrix, thus, first we need to construct the data. Then we simply copy the command as an input for the mentioned function and we will get a p-value for the null hypothesis that the marginal probabilities (n_{1+}, n_{+1}) are equal.

```
> matrix(c(9, 16, 37, 82), dimnames = list(c("Diabetes", "No Diabetes"),
+     c("Diabetes", "No Diabetes")), ncol = 2, byrow = TRUE)
```

```
            Diabetes No Diabetes
Diabetes           9          16
No Diabetes       37          82
```

```
> mcnemar.test(matrix(c(9, 16, 37, 82), ncol = 2, byrow = TRUE))
```

```
        McNemar's Chi-squared test with continuity correction

data:  matrix(c(9, 16, 37, 82), ncol = 2, byrow = TRUE)
McNemar's chi-squared = 7.5472, df = 1, p-value = 0.00601
```

The p-value provides evidence that the proportions of cases of Diabetes are different among people who suffered myocardial infarction and those who have not. The former is much grater than the latter.

Problem 2

As in the previous exercise, we first construct the matrix, to see if the data were correctly entered, but also to help us interpret the results.

```
> (tb.ex2 <- matrix(c(833, 125, 2, 160), ncol = 2, byrow = TRUE,
+     dimnames = list(Heaven = c("yes", "no"), Hell = c("yes",
+         "no"))))
```

```
      Hell
Heaven yes  no
   yes 833 125
   no    2 160
```

Note that the matrix was stored in a variable, that will be used later.

a) The McNemar test results in

```
> mcnemar.test(tb.ex2, correct = FALSE)
```

```
        McNemar's Chi-squared test

data:  tb.ex2
McNemar's chi-squared = 119.13, df = 1, p-value < 2.2e-16
```

There is strong evidence that the proportions people who believe in heaven and hell are different.

```
> tb.prop.ex2 <- prop.table(tb.ex2)
```

b) The differences of proportions is 0.11. Thus the proportion of people who believe in heaven is larger than those who believe in hell. Now the 90% confidence interval (CI) for this difference is calculated.

```
> tb.pdiff.ex2 <- margin.table(tb.prop.ex2, 1)[1] - margin.table(tb.prop.ex2,
+     2)[1]
> off.diag.ex2 <- diag(tb.prop.ex2[1:2, 2:1])
> tb.pdiff.ex2 + c(-1, 1) * qnorm(0.95) * sqrt((sum(off.diag.ex2) -
+     diff(off.diag.ex2)^2)/sum(tb.ex2))
```

```
[1] 0.09417584 0.12546701
```

Because the CI doesn't not include zero, we can be 90% sure that the difference of proportions is significant and there are between 0.09 and 0.12 more persons who believe in heaven.

Problem 3

a) According to Agresti (2007, p.248) the logit model for testing marginal homogeneity is:

$$P[Y_t = 1] = \alpha + \beta x_t$$

Where the parameter $exp(\beta)$ is the odds ratio comparing the marginal distributions and whose maximum likelihood estimate is the odds ratio of marginal distributions. Now for the previous table we calculate it with

```
> oddsratio(matrix(c(margin.table(tb.ex2, 1), margin.table(tb.ex2,
+     2)), ncol = 2), log = FALSE)
```

```
 odds ratios for and

[1] 2.018408
```

Thus, the odds of believing in heaven is twice that of believing in hell.

b) The conditional logit model is:

$$logitP[Y_{i1} = 1] = \alpha_i + \beta; logitP[Y_{i2} = 1] = \alpha_i$$

The ML estimate for $exp(\beta)$ is the quotient n_{12}/n_{21}, which, for the previous exercise, is $125/2 = 62.5$. So, conditioning on the subject, the odds of believing in heaven is 62.5 times the odd of believing in hell.

The same result can be calculated, as stated in Thompson (2007, see also the notes p. 4, option 1), with the function `clogit`. For this, we need to construct an ungrouped data with 3 columns, the first representing all the pairs for all the observations, in this case 1120 pairs; the second column is for the two observations (Heaven and Hell); and finally the third is for the number of individual partial tables, where 1 represents a “yes” response and 0 represents a “no”response, for example for the value 2 should be interpreted as a “no” response for the first observation and a “yes” response for the second observation (in R code `rep(c(0,1),2)`)

```
> tb.ex3 <- data.frame(pair = rep(1:1120, each = 2), Hev.Hell = rep(c(1,
+     0), 1120), Yes = c(rep(c(1, 1), 833), rep(c(1, 0), 125),
+     rep(c(0, 1), 2), rep(c(0, 0), 160)))
```

Then we can fit the data with the `clogit` function.

```
> (fit.clr.ex3 <- clogit(Yes ~ Hev.Hell + strata(pair), method = "exact",
+     data = tb.ex3))
```

```
Call:
clogit(Yes ~ Hev.Hell + strata(pair), method = "exact", data = tb.ex3)

           coef exp(coef) se(coef)     z        p
Hev.Hell  4.1352   62.5000   0.7127 5.802 6.56e-09

Likelihood ratio test=155.5  on 1 df, p=< 2.2e-16
n= 2240, number of events= 1793
```

The value `exp(coef)` represents the odds ratio of believing in heaven (1) relative to believing in hell (0).

Note about options It should be noted that the second option, described in Thompson (2007), results in a different odds ratio.

Problem 4

Both compare proportions assuming, as a null hypothesis, that they are equal. Now, as the enunciate says the *t test* assumes a normal distribution for the data, while the *McNemar test* assumes a binomial distribution.

Problem 5

a) As it's shown in section 8.1.1, the null hypothesis used by the McNemar test can be expressed as: $H_0 : \pi_{12} = \pi_{21}$, equality that is assumed to come from a binomial distribution with a probability of success (π_{12}) equal to 1/2. For the values reported in the table (132,107) the p-value under the null hypothesis can be calculated as follows

```
> 1 - pbinom(131, size = 239, prob = 0.5)
```

```
[1] 0.06018835
```

It should be noted that the pbinom function computes the lower tail of the probability distribution, given the data, *i.e.* all the equal or lower values than observed; however since we are interested in the other tail, we rest this value from 1 to calculate it.

Another way to calculate this value is with the exact2x2 function from the package with the same name. As explained in the package's help page, this function can compute a McNemar test with an alternative hypothesis (H_a) different from equality. For the case of $H_a : \pi_{12} > \pi_{21}$ we use the following code:

```
> (tb.ex5 <- matrix(c(227, 107, 132, 678), ncol = 2, dimnames = list(Taxes = c("yes",
+     "no"), Std = c("yes", "no"))))
```

```
     Std
Taxes yes  no
  yes 227 132
  no  107 678
```

```
> exact2x2(tb.ex5, conf.level = 0.95, alternative = "greater",
+     paired = TRUE)
```

```
        Exact McNemar-type test

data:  tb.ex5
b = 132, c = 107, p-value = 0.06019
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.9882151       Inf
sample estimates:
odds ratio
  1.233645
```

The paired=TRUE indicates that the data (tb.ex5, which is a matrix) is matched.

b)As it name suggest the mid-p-value uses only half of the p-value for the observed result and sums it to the p-values for more extreme results. Agresti (2007) discuss, in sections 1.4.4 and 1.4.5, that this is less conservative approach that will help us reject the null hypothesis with a p-value closer to the nominal one (*e.g.* α=0.05).
Now to compute this under R we can first compute the p-value for the observed data, in this case 132, divide it by 2 and then sum it to the p-value for more extreme data:

```
> # p-value for 132
> temp <- pbinom(132, size = 239, prob = 0.5) - pbinom(131, size = 239,
+     prob = 0.5)
> # p-value for values more extreme than 132 plus half the
> # observed p-value
> 1 - pbinom(132, size = 239, prob = 0.5) + temp/2
```

```
[1] 0.05319403
```

c) With an alternative hypothesis of inequality, the one sided-p-value gets doubled. This can be seen with the regular McNemar test:

```
> mcnemar.test(tb.ex5, correct = FALSE)
```

```
        McNemar's Chi-squared test

data:  tb.ex5
McNemar's chi-squared = 2.6151, df = 1, p-value = 0.1059
```

Now for the mid-p-value approach this value equals 0.1063881 which equals the regular p-value minus the p-value for the observed data. Thus, in the regular approach (McNemar test), the p-value for the observed data is counted twice.

Problem 6

First we construct the data. Note that in the following commands we first construct a data frame which latter will be transformed to a matrix for the stratum of interest.

```
> tb.ex6 <- data.frame(expand.grid(Info = c("yes", "no"), Gend = c("male",
+     "female"), Op = c("yes", "no")), C = c(76, 6, 114, 11, 160,
+     25, 181, 48))
```

a) The equality of proportions can be calculated with the `mcnemar.exact` function from the `exact2x2` package, which has the advantage of also computing the confidence interval.

```
> mcnemar.exact(xtabs(C ~ Info + Op + Gend, data = tb.ex6)[, ,
+     2], conf.level = 0.9)
```

```
        Exact McNemar test (with central confidence intervals)

data:  xtabs(C ~ Info + Op + Gend, data = tb.ex6)[, , 2]
b = 181, c = 11, p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
90 percent confidence interval:
  9.746948 29.809005
sample estimates:
odds ratio
  16.45455
```

The p-value provides strong evidence against equality of proportions, just as the odds ratio for being in favor of the first option, which is 16 times the odds of the second.

b) The confidence interval for the difference of proportions is computed as in Thompson (2007).

```
> tb.fem.ex6 <- xtabs(C ~ Info + Op + Gend, data = tb.ex6)[, ,
+     2]
> tb.prop.ex6 <- prop.table(tb.fem.ex6)
> tb.pdiff.ex6 <- margin.table(tb.prop.ex6, 1)[1] - margin.table(tb.prop.ex6,
+     2)[1]
> off.diag.ex6 <- diag(tb.prop.ex6[1:2, 2:1])
> tb.pdiff.ex6 + c(-1, 1) * qnorm(0.95) * sqrt((sum(off.diag.ex6) -
+     diff(off.diag.ex6)^2)/sum(tb.fem.ex6))
```

```
[1] 0.4314133 0.5290387
```

This means that the difference of proportions among those who are in favor of the first option and those that are in favor of the second is at least 0.43 and at most 0.53.

c) For the marginal model, the odds ratio is calculated as explained in Agresti (2007, p. 248). We can calculate it as it was done in problem 3 or we can define a function (to bypass the need for the `vcd` package)

```
> marg.logit <- function(x) {
+     (margin.table(x, 1)[1]/margin.table(x, 1)[2])/(margin.table(x,
+         2)[1]/margin.table(x, 2)[2])
+ }
> marg.logit(tb.fem.ex6)
```

```
 yes
9.16
```

According to this, the odds of being in favor of the information option is 9 times those in favor of the government paying for all the cost.

For the conditional model we use the `clogit` function as in problem 3. Remember that we need to construct an ungrouped table, first

```
> tb2.ex6 <- data.frame(pair = rep(1:354, each = 2), Info = rep(c(1,
+     0), 354), Op = c(rep(c(1, 1), 114), rep(c(1, 0), 181), rep(c(0,
+     1), 11), rep(c(0, 0), 48)))
> (fit.CLR.ex6 <- clogit(Op ~ Info + strata(pair), method = "exact",
+     data = tb2.ex6))
```

```
Call:
clogit(Op ~ Info + strata(pair), method = "exact", data = tb2.ex6)

       coef exp(coef) se(coef)     z      p
Info  2.8006   16.4545   0.3105 9.019 <2e-16
```

```
Likelihood ratio test=181.9  on 1 df, p=< 2.2e-16
n= 708, number of events= 420
```

According to the conditional model, the odds of being in favor for the information approach to deal with AIDS is 16 times those of being in favor of paying all the expenses.

d) Because the gender samples are independent, we can compare whether the difference of proportions for each of the options is dependent or not on gender, with a standard test for comparing two proportions such as the `prop.test` function. Remember that the null hypothesis is that proportions are independent.

```
> prop.test(xtabs(C ~ Gend + Info, data = tb.ex6), correct = FALSE,
+     conf.level = 0.9)
```

```
        2-sample test for equality of proportions without continuity
        correction

data:  xtabs(C ~ Gend + Info, data = tb.ex6)
X-squared = 3.1399, df = 1, p-value = 0.0764
alternative hypothesis: two.sided
90 percent confidence interval:
 0.004720727 0.096402869
sample estimates:
   prop 1    prop 2
0.8838951 0.8333333
```

```
> prop.test(xtabs(C ~ Gend + Op, data = tb.ex6), correct = FALSE,
+     conf.level = 0.9)
```

```
        2-sample test for equality of proportions without continuity
        correction

data:  xtabs(C ~ Gend + Op, data = tb.ex6)
X-squared = 1.4487, df = 1, p-value = 0.2287
alternative hypothesis: two.sided
90 percent confidence interval:
 -0.1084578  0.0164753
sample estimates:
   prop 1    prop 2
0.3071161 0.3531073
```

For both options the p-value indicates independence, however for the comparison between gender and the information approach the 90% CI does not contain the zero value, contradicting the p-value. Note that the `xtabs` function was used on the first data frame in order to transform the data.

Problem 7

First we construct the data, which is a table of independent samples. Then we can use the `prop.test` function to test the null hypothesis that the “yes” probabilities are equal.

```
> (tb.ex7 <- matrix(c(359, 785, 334, 810), byrow = TRUE, ncol = 2,
+     dimnames = list(c("High taxes", "Cut stand."), c("yes", "no"))))
```

```
           yes  no
High taxes 359 785
Cut stand. 334 810
```

```
> prop.test(tb.ex7, correct = FALSE)
```

```
        2-sample test for equality of proportions without continuity
        correction

data:  tb.ex7
X-squared = 1.2937, df = 1, p-value = 0.2554
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01579280  0.05949909
sample estimates:
   prop 1    prop 2 
0.3138112 0.2919580 
```

The p-value (0.25) indicate that we can not reject the null hypothesis of equality of proportions. This is supported by the CI which includes the zero value. Now, comparing this result with that obtained in section 8.1.2 we note that the width of the CI is wider for the independence test (-0.016,0.059) compared with the one calculated for the matched pair (-0.004,0.048). At the end of section 8.1.2 Agresti (2007) mentions that, the estimated variance used for matched data considers the dependence of the marginal proportions through their covariance, which translates in a lower value and finally in a narrower confidence interval.

Problem 8

a) First we construct the data, then we apply the `mcnemar.exact` function to compute the McNemar test.

```
> (tb1.ex8 <- matrix(c(16, 22, 45, 17), ncol = 2, dimnames = list(A = c("+",
+     "-"), B = c("+", "-"))))
```

```
   B
A    +  -
  + 16 45
  - 22 17
```

```
> library(exact2x2)
> mcnemar.exact(tb1.ex8)
```

```
        Exact McNemar test (with central confidence intervals)
```

```
data:  tb1.ex8
b = 45, c = 22, p-value = 0.006741
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.202879 3.576800
sample estimates:
odds ratio
  2.045455
```

The p-value indicates that there isn't marginal homogeneity. This is supported by the odds ratio that indicates that the odds of success for drug A is twice that for drug B, 95% CI (1.203,3.577).

b) Table 8.11 is a table for 2 independent samples, the two groups that took the drugs in different order. For this type of data the independence test has the null hypothesis that $H_0 : \pi_1 = \pi_2$ which can be interpreted as: the probability of success of drug A (π_1) equals that for drug B (π_2). In R we compute this as follows.

```
> (tb2.ex8 <- matrix(c(25, 12, 10, 20), ncol = 2, dimnames = list(c("A-B",
+     "B-A"), c("First", "Second"))))
```

```
    First Second
A-B    25     10
B-A    12     20
```

```
> prop.test(tb2.ex8, correct = FALSE)
```

```
        2-sample test for equality of proportions without continuity
        correction

data:  tb2.ex8
X-squared = 7.7822, df = 1, p-value = 0.005276
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1144861 0.5640853
sample estimates:
   prop 1    prop 2
0.7142857 0.3750000
```

The p-value rejects the null hypothesis, as well as the CI for the difference of proportions, which doesn't include the zero value (0.114,0.564) and, just as the previous test, indicates that the probability of success for drug A is higher than that for B.

Problem 9

First we construct an ungrouped data frame, as it is described in the notes, section 3.2.1, then we fit the model with the `clogit` function.

```
> tb.ex9 <- data.frame(pair = rep(1:1144, each = 2), Cut = rep(c(0,
+     1), 1144), Yes = c(rep(c(1, 1), 227), rep(c(1, 0), 132),
+     rep(c(0, 1), 107), rep(c(0, 0), 678)))
> (fit.CLR.ex9 <- clogit(Cut ~ Yes + strata(pair), method = "exact",
+     data = tb.ex9))
```

```
Call:
clogit(Cut ~ Yes + strata(pair), method = "exact", data = tb.ex9)

       coef exp(coef) se(coef)      z     p
Yes -0.2100    0.8106   0.1301 -1.614 0.106

Likelihood ratio test=2.62  on 1 df, p=0.1055
n= 2288, number of events= 1144
```

The results indicate that, in order to help the environment, the odds ratio of being in favor of a cut in living standards is 19% lower than the odds of being in favor of a raise in taxes. Changing the sign of the estimate gives us the reverse odd, *i.e.* the one for being in favor of a raise in taxes rather than a cut in living expenses. This last value would be the same as the one in Agresti (2007, p. 250), however there, it was calculated by dividing the cells n_{12} and n_{21}.

Note about the table The table constructed in the above command, should be interpreted as follows. First, the `pair` factor is one that indicates the pairs of data. Second, the `Cut` factor represents the "response" variable ($0 =$ negarive response, $1 =$ possitive response), in other words the logit of being in favor of that option instead of the other. Third the `Yes` factor represents the cells of the table repeated count times; for this the first number represent the answer for the "predictor" variable (1 or 0) and the second number represent the answer for the "response" variable. For example the argument, `rep(c(1,0),107` represents the cell for those persons who are in favor of a raise in taxes and are against a cut in living standards; this is considering that we considered the cut of living standards as a "response" variable.
It should be noted that when fitting the conditional logistic model, it doesn't matter which variable is specified as a response variable in the formula, the result is the same. This could lead to confusions at the moment of interpretating the results, but (I think) if one follows the logit of the previous paragraph there would be less confusion. Take this into consideration while interpreting problem 28

Problem 10

a) The population averaged table and the subject specific tables are the following:
b) In R the two types of table are constructed with:

```
> # Population-averaged
> tb1.ex10 <- matrix(c(1, 1, 3, 3), ncol = 2, dimnames = list(Control = c("low",
+     "high"), Case = c("low", "high")))
> # Subject-specific
> tb2.ex10 <- data.frame(pair = rep(1:8, each = 2), Subject = rep(c("control",
+     "case"), 8), Meat = c(c("low", "low"), rep(c("high", "high"),
+     3), rep(c("low", "high"), 3), c("high", "low")))
> tb2.ex10$Meat <- factor(tb2.ex10$Meat, levels = c("low", "high"))
> tb2.ex10$Subject <- factor(tb2.ex10$Subject, levels = c("control",
+     "case"))
```

Control	Case	
	low	high
low	1	3
high	1	3

Subject	Meat	
	low	high
Control	1	0
Case	1	0

Subject	Meat	
	low	high
Control	1	0
Case	0	1

Subject	Meat	
	low	high
Control	0	1
Case	1	0

Subject	Meat	
	low	high
Control	0	1
Case	0	1

Table 1: Population (first row) and subject specific (second row) tables for exercise 10. For the subject specific tables the data suggest that there are, in order, 1, 3, 1, 3 tables

Note that the subject-specific table is what we've been calling the "ungrouped" data.

Once we got the data we can perform the two tests.

```
> mcnemar.test(tb1.ex10, correct = FALSE)
```

```
        McNemar's Chi-squared test

data:  tb1.ex10
McNemar's chi-squared = 1, df = 1, p-value = 0.3173
```

```
> mantelhaen.test(xtabs(~Subject + Meat + pair, data = tb2.ex10),
+     correct = FALSE)
```

```
        Mantel-Haenszel chi-squared test without continuity correction

data:  xtabs(~Subject + Meat + pair, data = tb2.ex10)
Mantel-Haenszel X-squared = 1, df = 1, p-value = 0.3173
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
  0.3120602 28.8405896
sample estimates:
common odds ratio
                3
```

Both test yield the same result, there is marginal homogeneity (p-value = 0.3173). This confirms what it's said in Agresti (2007, section 8.2.5), that "the McNemar test is a special case of the Cochran-Mantel-Haenszel (CMH) test applied to the binary responses of n matched pairs displayed in n partial tables.

c) To remove those subjects with the same diet we use the function `subset` and use its result as the `data` argument for the previous command:

```
> mantelhaen.test(xtabs(~Subject + Meat + pair, data = subset(tb2.ex10,
+     pair != 1)), correct = FALSE)
```

```
        Mantel-Haenszel chi-squared test without continuity correction

data:  xtabs(~Subject + Meat + pair, data = subset(tb2.ex10, pair != 1))
Mantel-Haenszel X-squared = 1, df = 1, p-value = 0.3173
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
  0.3120602 28.8405896
sample estimates:
common odds ratio
                3
```

```
> mantelhaen.test(xtabs(~Subject + Meat + pair, data = subset(tb2.ex10,
+     pair != 1, pair != 2 & pair != 3 & pair != 4)), correct = FALSE)
```

```
        Mantel-Haenszel chi-squared test without continuity correction

data:  xtabs(~Subject + Meat + pair, data = subset(tb2.ex10, pair != 1, pair != 2 & pair != 3 & pa
Mantel-Haenszel X-squared = 1, df = 1, p-value = 0.3173
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
  0.3120602 28.8405896
sample estimates:
common odds ratio
                3
```

Note that whether we remove just one pair (like in the first command, where the pair with low meat consumption where removed) or when we remove all the pairs with the same diet (second command), the p-value doesn't differ.

d) As it is explained in exercise 5, there are two ways of calculating the exact p-value 2-way matched tables, the first one is done assuming a binomial distribution for n_{12} and n_{21}, while the second one is through the `exact2x2` function. Note, however that this point ask for a one-sided alternative, while the previous ones calculated a two-sided alternative.

```
> 1 - pbinom(2, size = 4, prob = 0.5)
```

```
[1] 0.3125
```

```
> exact2x2(t(tb1.ex10), paired = TRUE, alternative = "greater")
```

```
        Exact McNemar-type test

data:  t(tb1.ex10)
b = 1, c = 3, p-value = 0.9375
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.01290589        Inf
sample estimates:
odds ratio
 0.3333333
```

In the exercise 5 it was explained that the pbinom function calculates the lower tail, so to compute the equal or more extreme results we need to rest the value from 1.

Problem 11

Agresti (2007, p. 269) explains that the estimated variance (σ^2) for the estimate of β is $1/n_{12}+1/n_{12}$, in this case this value equals $1/132+1/107=0.01692155$, thus the standard error is $\sqrt{0.01692155} = 0.1300829$. With this value we can calculate the 95% confidence interval as follows (see also the notes, section 2.1.1)

```
> CI.ex11 <- 0.21 + c(-1, 1) * qnorm(0.975) * 0.1300829
> round(exp(CI.ex11), 3)
```

```
[1] 0.956 1.592
```

Note that we need to exponentiate the boundaries to get the CI for the odds ratio, otherwise it will be a CI for the log odds ratio.
In R, the function clogit also calculates the confidence interval for the odds ratio, so to see it we need to recall the variable (fit.CLR.ex9) calculated in exercise 9.

```
> summary(fit.CLR.ex9)
```

```
Call:
coxph(formula = Surv(rep(1, 2288L), Cut) ~ Yes + strata(pair),
    data = tb.ex9, method = "exact")

  n= 2288, number of events= 1144

       coef exp(coef) se(coef)      z Pr(>|z|)
Yes -0.2100    0.8106   0.1301 -1.614    0.106

    exp(coef) exp(-coef) lower .95 upper .95
Yes    0.8106      1.234    0.6282     1.046

Concordance= 0.511  (se = 0.01 )
Likelihood ratio test= 2.62  on 1 df,   p=0.1
Wald test            = 2.61  on 1 df,   p=0.1
Score (logrank) test = 2.62  on 1 df,   p=0.1
```

This interval can be interpreted as: we are 95% confident that the true odds ratio between approval for a raise in taxes relative to a cut in living standards is at least 4% lower than the latter and at most 60% higher than the latter.

Problem 12

```
> tb.ex12 <- data.frame(expand.grid(cigarrette=c('yes', 'no'),
+                                    alcohol=c('yes', 'no'),  marijuana=c('yes', 'no')),
+                        count=c(911, 44, 3, 2, 538, 456, 43, 279))
> tb.marijuana.ex12 <- xtabs(count ~ alcohol + cigarrette, data = tb.ex12)
> tb.alcohol.ex12 <- xtabs(count ~ alcohol + cigarrette, data = tb.ex12)
> tb.cigarrette.ex12 <- xtabs(count ~ cigarrette + alcohol, data = tb.ex12)
```

Finally, with these tables we can calculate the proportion of consumption of each substance. Note that, because any individual can consume more than one substance, the proportions do not sum to 1.

```
> round(margin.table(prop.table(tb.alcohol.ex12), 1)[2], 3)
```

```
   no
0.144
```

```
> round(margin.table(prop.table(tb.cigarrette.ex12), 1)[2], 3)
```

```
   no
0.343
```

```
> round(margin.table(prop.table(tb.marijuana.ex12), 1)[2], 3)
```

```
   no
0.144
```

b) As it is explained in Agresti (2007, sections 2.7.6 and 4.3.4), the CMH test evaluates the null hypothesis of homogeneous association, *i.e.* that the odds ratio among two factors is the same across a third one. More over, when this is true, homogeneous association also exist for any pair formed. Usually the third factor is a control variable, for example treatment center in the study of success of two drugs; now, in this case there is no clear distinction about which factor can be considered a "control", so we can choose any one.

```
> mantelhaen.test(xtabs(count ~ marijuana + cigarrette + alcohol,
+     data = tb.ex12), correct = FALSE)
```

```
        Mantel-Haenszel chi-squared test without continuity correction

data:  xtabs(count ~ marijuana + cigarrette + alcohol, data = tb.ex12)
Mantel-Haenszel X-squared = 441.83, df = 1, p-value < 2.2e-16
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 12.58828 24.00445
sample estimates:
common odds ratio
         17.38317
```

Problem 13

Just as it is explained in the notes (section 5.1.1), we need to construct the data frame, and append a column that represents the values of the main diagonal and above elements. This is done with the `pmin` and `pmax` functions:

```
> tb.ex13 <- data.frame(expand.grid(Affil.1 = c(1:4),
+     Affil.2 = c(1:4)), C = rev(c(1228,
+     100, 1, 73, 39, 649, 0, 12, 2, 1, 54, 4, 158, 107, 9, 137)))
> tb.ex13$symm <- paste(pmin(tb.ex13$Affil.1, tb.ex13$Affil.2),
+     pmax(tb.ex13$Affil.1, tb.ex13$Affil.2), sep = ",")
> tb.ex13$symm <- factor(tb.ex13$symm)
> levels(tb.ex13$symm) <- rev(levels(tb.ex13$symm))
```

It's known that in R the first level is set to zero, so while constructing the data we specified the factor ordering. Because of this, the ordering of the character vector symm is also reversed.
With the data constructed we can fit the model and then calculate the residuals. For the latter point, note that for a easier interpretation the residuals are expressed as a matrix.

```
> fit.symm.ex13 <- glm(C ~ symm, family = poisson(link = log),
+     data = tb.ex13)
> res.ex13 <- resid(fit.symm.ex13, type = "pearson")/
+     sqrt(1 - lm.influence(fit.symm.ex13)$hat)
> xtabs(res.ex13 ~ Affil.1 + Affil.2, data = tb.ex13)
```

```
       Affil.2
Affil.1          1          2          3          4
      1       -Inf -1.3867505 -8.7086357 -5.5925894
      2  1.3867505       -Inf -1.0000000 -0.5773503
      3  8.7086357  1.0000000       -Inf  5.1739525
      4  5.5925894  0.5773503 -5.1739525        Inf
```

Agresti (2007) explains that the residuals for the symmetry model equal:

$$r_{ij} = (n_{ij} - n_{ji})/sqrtn_{ij} + n_{ji}$$

When $i = j$ the residual isn't defined (0/0) therefore in the matrix is expressed as Inf. For the rest of the residuals we can say that there are more changes, than what would be expected under the symmetry model, from categories 2 to 4, 2 to 1 and 1 to 4. Interesting two of these residuals represent a change to another religion (other than the common ones) or even a disbelief of religion.

b) To fit a quasi-symmetry model we need to add another character vector to the data frame:

```
> tb.ex13$Affil.1a <- factor(tb.ex13$Affil.1)
> (fit.qsymm.ex13 <- glm(C ~ symm + Affil.1a, family = poisson(link = log),
+     data = tb.ex13))
```

```
Call:  glm(formula = C ~ symm + Affil.1a, family = poisson(link = log),
    data = tb.ex13)

Coefficients:
(Intercept)      symm3,4      symm3,3      symm2,4      symm2,3
     4.9200      -3.4061      -2.1957      -0.6828      -1.5520
    symm2,2      symm1,4      symm1,3      symm1,2      symm1,1
```

```
   -7.0794      -5.2561      -0.3623      -2.1985       1.3440
 Affil.1a2    Affil.1a3    Affil.1a4
    0.6210       1.9177       0.8492

Degrees of Freedom: 15 Total (i.e. Null);  3 Residual
Null Deviance:	    6045
Residual Deviance: 2.32 	AIC: 107.9
```

This model has a deviance of 2.32 on 3 degrees of freedom, which has a p-value of 0.491 for the null hypothesis that any other parameter that is not included in the model equals zero; in other words that the model fits well.

c) With those models already fitted we can compare them to see if there is *marginal homogeneity*. For this, remember that the symmetry model is a special case of the quasi-symmetry model, therefore we can test this with anova function

```
> anova(fit.symm.ex13, fit.qsymm.ex13, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: C ~ symm
Model 2: C ~ symm + Affil.1a
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         6     150.59
2         3       2.32  3   148.27 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value provides extreme evidence against the null hypothesis of marginal homogeneity.

Problem 14

This problem is similar to the previous one, so the same approach should be used, however to save some time we can construct the complete data frame (*i.e.* one with all the character columns) at once.

```
> tb.ex14 <- data.frame(expand.grid(Resid.1 = factor(c("Northeast",
+     "Midwest", "South", "West"), levels = rev(c("Northeast",
+     "Midwest", "South", "West"))), Resid.2 = factor(c("Northeast",
+     "Midwest", "South", "West"), levels = rev(c("Northeast",
+     "Midwest", "South", "West")))), C = c(425, 10, 7, 5, 17,
+     555, 34, 14, 80, 74, 771, 29, 36, 47, 33, 452))
> tb.ex14$symm <- paste(pmin(as.numeric(tb.ex14$Resid.1),
+                            as.numeric(tb.ex14$Resid.2)),
+     pmax(as.numeric(tb.ex14$Resid.1), as.numeric(tb.ex14$Resid.2),
+         sep = ","))
> tb.ex14$symm <- factor(tb.ex14$symm, levels = rev(unique(tb.ex14$symm)))
> tb.ex14$Resid.1a <- factor(tb.ex14$Resid.1, levels = unique(tb.ex14$Resid.1))
```

a) First we fit the symmetry model and compute its residuals to see if they aren't too extreme (above 2 or even 3) under the null hypothesis.

```
> fit.symm.ex14 <- glm(C ~ symm, family = poisson(link = log),
+     data = tb.ex14)
> xtabs(resid(fit.symm.ex14, type = "pearson")/
+       sqrt(1 - lm.influence(fit.symm.ex14)$hat) ~
+     Resid.1 + Resid.2, data = tb.ex14)
```

```
           Resid.2
Resid.1           West      South    Midwest  Northeast
  West            -Inf -0.5080005 -4.2252170 -4.8413866
  South      0.5080005       -Inf -3.8490018 -7.8264215
  Midwest    4.2252170  3.8490018       -Inf -1.3471506
  Northeast  4.8413866  7.8264215  1.3471506       -Inf
```

Note that migrations from the Midwest or northeast regions to either west or south regions are much higher than the model allows, so the model doesn't seem to fit well.
Now we fit the quasi-symmetry model and calculate the standardized residuals.

```
> fit.qsymm.ex14 <- glm(C ~ symm + Resid.1a, family = poisson(link = log),
+     data = tb.ex14)
> xtabs(resid(fit.qsymm.ex14, type = "pearson")/
+       sqrt(1 - lm.influence(fit.qsymm.ex14)$hat) ~
+     Resid.1 + Resid.2, data = tb.ex14)
```

```
           Resid.2
Resid.1           West      South    Midwest  Northeast
  West            -Inf  0.1730133 -0.4640190  0.4072944
  South     -0.1730133        Inf  1.5261332 -1.7556050
  Midwest    0.4640190 -1.5261332        Inf  1.5675708
  Northeast -0.4072941  1.7556039 -1.5675697        Inf
```

Note than under this model, no residual is higher than 2, an indication that the model fits well. More over, the p-value for the deviance (3.932 on 3 df) is 0.731.

b) With those models the null hypothesis of marginal homogeneity is tested as follows:

```
> anova(fit.symm.ex14, fit.qsymm.ex14, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: C ~ symm
Model 2: C ~ symm + Resid.1a
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         6    134.452
2         3      3.932  3   130.52 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note the extremely low p-value for the null hypothesis.

Note Agresti (2007, section 8.4.1) indicates that the symmetry model rarely fits real data well, which is the case of the previous two exercises.

Problem 15

To calculate the value of β, and its standard error, in the model,

$$logit[P(Y_{i1} \leq j)] = \alpha_{ij} + \beta \qquad logit[P(Y_{i2} \leq j)] = \alpha_{ij}$$

We can use equation 8.7 from Agresti, which in R are introduced as follow,

Then we use this function introducing table tb.ex15 but in the form of a matrix, this is with the use of the xtabs function.

```
> tb.ex15 <- data.frame(expand.grid(Pre.sex=c(1:4), Post.sex=c(1:4)),
+                        C=c(144, 33, 84, 126, 2, 4, 14, 29, 0, 2, 6, 25, 0, 0, 1, 5))
> tb.ex15$symm <- paste( pmin(as.numeric(tb.ex15$Pre.sex),
+                            as.numeric(tb.ex15$Post.sex)),
+                       pmax(as.numeric(tb.ex15$Pre.sex),
+                            as.numeric(tb.ex15$Post.sex)), sep=",")
> tb.ex15$symm <- factor(tb.ex15$symm, levels=rev(unique(tb.ex15$symm)))
> tb.ex15$scores <- rep(1:4, each=4)
> ord.beta.se(xtabs(C ~ Pre.sex + Post.sex, data = tb.ex15))
```

```
[1] -4.906755
[1] 0.4512518
```

The first value is the estimate and the second its standard error.

Thompson (2007, p. 190) describes an easier method to calculate β (denominated Tau, τ) and its standard error, but the results are different from what is found with formula 8.7.
To do this we need to create a vector that corresponds to a matrix with all the above-diagonal elements set to one and the rest set to zero. Then we fit the symmetry model adding this vector to the formula.

```
> temp <- matrix(0, nr = 4, nc = 4)
> tau <- as.numeric(row(temp) < col(temp))
> summary(glm(C ~ symm + tau, family = poisson(log), data = tb.ex15))$coefficients
```

```
              Estimate Std. Error    z value     Pr(>|z|)
(Intercept)  1.6094379  0.4472136  3.5988126 3.196735e-04
symm3,4      1.6327093  0.4883773  3.3431310 8.283879e-04
symm3,3      0.1823216  0.6055301  0.3010941 7.633427e-01
symm2,4      1.7419086  0.4842867  3.5968540 3.220891e-04
symm2,3      1.1472015  0.5123972  2.2388912 2.516299e-02
symm2,2     -0.2231436  0.6708204 -0.3326428 7.394040e-01
symm1,4      3.2108947  0.4560563  7.0405659 1.914606e-12
symm1,3      2.8054296  0.4603864  6.0936413 1.103708e-09
symm1,2      1.9299608  0.4781446  4.0363537 5.428836e-05
symm1,1      3.3603754  0.4549115  7.3868777 1.503167e-13
tau         -4.1303550  0.4507872 -9.1625374 5.069112e-20
```

Note the differences in the values for the estimates (-4.91 vs -4.13) however their standard errors are basically the same (0.4512 vs 0.4508).

a) With the former values we need to divide the estimate by its standard error to find the *z-value* which is -10.87365 with a p-value of 7.695707e-28 for the null hypothesis of $H_0 : \beta = 0$. This provides strong evidence agains marginal homogeneity. Now the value of β can be interpreted as the estimated probability that the first scenario (premarital sex) is closer to the lower end of the scale (in this case considered more wrong) than the second scenario (post marital sex). In this case this is exp(-4.91)=0.007 times the estimated probability that extramarital sex is considered more wrong than pre-marital sex.

b) This was already done, the *z-value* is -10.87365 with a p-value of 7.695707e-28 for the null hypothesis of $H_0 : \beta = 0$.

c) Since the symmetry model is a special case of the quasi-symmetry model, we can compare both models, with `anova` function, to test the hypothesis of marginal homogeneity.

```
> fit.symm.ex15 <- glm(C ~ symm, family = poisson(link = log),
+     data = tb.ex15)
> fit.qsymm.ex15 <- glm(C ~ symm + scores, family = poisson(link = log),
+     data = tb.ex15)
> anova(fit.symm.ex15, fit.qsymm.ex15, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: C ~ symm
Model 2: C ~ symm + scores
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         6      402.2
2         5        2.1  1   400.11 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The estimate for the score's levels can be interpreted as, that the estimated odds ratio of judging post-marital sex in one of the first three categories to category 4 “not wrong at all” compared to the similar odds for the pre-marital sex. For example the odds ratio of category 2 (“almost always wrong”) to 4 (“not wrong at all”) for the post-marital sex is $exp(5.642) = 282.03$ times the odds of judging pre-marital sex as “almost always wrong” to “not wrong at all”.
The p-value of the variance comparison provides extremely strong evidence against marginal homogeneity.

d) To fit an ordinal quasi-symmetry model we need to include a vector of scores that is treated as `numeric`. Then we do the same thing as for the previous part.

```
> fit.oqsymm.ex15 <- glm(C ~ symm + scores, family = poisson(link = log),
+     data = tb.ex15)
> anova(fit.symm.ex15, fit.oqsymm.ex15, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: C ~ symm
Model 2: C ~ symm + scores
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         6      402.2
2         5        2.1  1   400.11 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The estimate for the scores' effect can be interpreted as that the estimated probability that premarital sex is judged n categories away from post marital sex is $exp(n \times -2.86)$, for example the odds of judging pre-marital sex as “always wrong” (1) instead of judging post-marital sex as “wrong only sometimes” (3) is $exp(1 - 3 \times -2.857) = 303.081$ times the reverse, namely, that pre-marital sex is judged as “always wrong” and post-marital sex is judged as “wrong only sometimes”. Again this shows strong evidence against marginal homogeneity.

e) That is what the sign of β indicates, by beeing negative the odds of being in the first categories of the response (*i.e.* towards a wrong judgement) diminish for the first scenario $logit[P(Y_{i1} \leq j)] = \alpha_{ij} + \beta$ which is pre-marital sex.

Problem 16

As it's explained in Thompson (2007) in order to fit an ordinal quasi-symmetry model we need to create a column of scores in addition to the symmetry column (`symm`). Now for the quasi-symmetry model we need an extra column that is nothing but the first factor converted to a character vector. Thus we construct the table as follows.

```
> tb.ex16 <- data.frame(expand.grid(Chem.free = 1:3, Recy = 1:3),
+     C = c(66, 227, 150, 39, 359, 216, 3, 48, 108))
> tb.ex16$symm <- paste(pmin(tb.ex16$Chem.free, tb.ex16$Recy),
+     pmax(tb.ex16$Chem.free, tb.ex16$Recy), sep = ",")
> tb.ex16$symm <- factor(tb.ex16$symm, levels = rev(unique(tb.ex16$symm)))
> tb.ex16$Recy1 <- factor(tb.ex16$Recy, levels = rev(unique(tb.ex16$Recy)))
> tb.ex16$score <- tb.ex16$Recy
```

Then we can fit the models:

```
> fit.symm.ex16 <- glm(C ~ symm, family = poisson(log), data = tb.ex16)
> fit.qsymm.ex16 <- glm(C ~ symm + Recy1, family = poisson(log),
+     data = tb.ex16)
> fit.oqsymm.ex16 <- glm(C ~ symm + score, family = poisson(log),
+     data = tb.ex16)
```

Whose residuals are: Comparing these tables it's clear that the symmetry model doesn't hold (Deviance = 445.23). Now comparing the two other models, the residuals for the ordinal quasi-symmetry model seem to indicate that this is the best one, however its deviance value isn't the lowest, 2.47 versus 1.23 for the quasi-symmetry model.

Problem 17

By looking at the table we can see that there are two models that, in principle, can fit the data, the *quasi-symmetry* and the *ordinal quasi-symmetry* models. Thus, to fit this we construct the data frame with the necessary variables.

	1	2	3
1	-Inf	-11.53	-11.88
2	11.53	Inf	-10.34
3	11.88	10.34	-Inf

	1	2	3
1	-Inf	1.04	-1.04
2	-1.04	-Inf	1.04
3	1.04	-1.04	-Inf

	1	2	3
1	-Inf	-0.56	-1.06
2	0.56	Inf	1.46
3	1.06	-1.46	-Inf

Table 2: Residuals from the symmetry, quasi-symmetry and ordinal quasi-symmetry models for the data of table 8.15

	Chem.free	Recy	C	Symm	Q-symm	O-qsymm
1	1	1	66.00	66.00	66.00	66.00
2	2	1	227.00	133.00	229.00	224.40
3	3	1	150.00	76.50	148.00	147.90
4	1	2	39.00	133.00	37.00	41.60
5	2	2	359.00	359.00	359.00	359.00
6	3	2	216.00	132.00	218.00	222.70
7	1	3	3.00	76.50	5.00	5.10
8	2	3	48.00	132.00	46.00	41.30
9	3	3	108.00	108.00	108.00	108.00

```
> tb.ex17 <- data.frame(expand.grid(Car.p = 1:4, G.effect = 1:4),
+     C = c(95, 66, 31, 5, 72, 129, 101, 4, 32, 116, 233, 24, 8,
+         13, 82, 26))
> tb.ex17$symm <- paste(pmin(tb.ex17$Car.p, tb.ex17$G.effect),
+     pmax(tb.ex17$Car.p, tb.ex17$G.effect), sep = ",")
> tb.ex17$symm <- factor(tb.ex17$symm, levels = rev(unique(tb.ex17$symm)))
> tb.ex17$scores <- tb.ex17$G.effect
> tb.ex17$Car.pf <- factor(tb.ex17$Car.p, levels = rev(unique(tb.ex17$Car.p)))
```

Then we can fit the two models

```
> fit.qsymm.ex17 <- glm(C ~ symm + Car.pf, family = poisson(log),
+     data = tb.ex17)
> fit.oqsymm.ex17 <- glm(C ~ symm + scores, family = poisson(log),
+     data = tb.ex17)
```

With these models we can calculate the standardized residuals Note that the residuals for the

	Greenhouse effect			
	1	2	3	4
1	Inf	0.90	-0.30	-1.40
2	-0.90	Inf	0.90	0.00
3	0.30	-0.90	Inf	1.00
4	1.40	0.00	-1.00	Inf

	Greenhouse effect			
	1	2	3	4
1	Inf	-1.00	-2.10	-0.50
2	1.00	Inf	-0.90	1.30
3	2.10	0.90	Inf	4.70
4	0.50	-1.30	-4.70	Inf

Table 3: Standardized residuals for the quasi-symmetry (left) and ordinal quasi-symmetry models (right). The response categories are: 1 = extremely dangerous, 2 = very dangerous, 3 = somewhat dangerous, 4 = not or not very dangerous.

quasi-symmetry model are all below 1.5, however those from the ordinal quasi-symmetry model are not bad either, except for those who consider car pollution as “somewhat dangerous” and also consider greenhouse effect as “not dangerous at all” which are much higher that what the models allows.
We can also compare the deviance values for the two models which are 2.23 for the quasi-symmetry model and 27.49 for the ordinal quasi-symmetry model. Despite the small differences in the residuals the deviance values are very different; According to the p-value for the deviance values the ordinal quasi-symmetry model isn’t correct (p-value = 0).

Assuming that the quasi-symmetry model holds, we can calculate the odds ratio for any π_{ij}/π_{ji} combination. For this we would use the coefficients for the `Car.pf` factor. For example the odds for a person to believe that the car pollution is “very dangerous” (i=2) and the greenhouse effect is “somewhat dangerous” (j=3) instead of believing the opposite (*i.e.* car pollution is somewhat dangerous and greenhouse effect is very dangerous) is exp(1.204-1.120) = 1.088. The values are not particularly big except when comparing any category with the last one.

Coefficient	Value
Car.pf-1	1.207
Car.pf-2	1.204
Car.pf-3	1.120

Table 4: Coefficients used for the calculation of the odds ratio

Problem 18

Because the data is based on ordinal factors we can fit a ordinal quasi-symmetry model as well as a quasi-symmetry model.
First the data. Note that we need to construct two data frames, one for the control group and one for the treatment group.

```
> tb.c.ex18 <- data.frame(expand.grid(begin = c("<3.4", "3.4-4.1",
+     "4.1-4.9", ">4.9"), end = c("<3.4", "3.4-4.1", "4.1-4.9",
+     ">4.9")), C = c(18, 16, 0, 0, 8, 30, 14, 2, 0, 13, 28, 15,
+     0, 2, 7, 22))
> tb.t.ex18 <- data.frame(expand.grid(begin = c("<3.4", "3.4-4.1",
+     "4.1-4.9", ">4.9"), end = c("<3.4", "3.4-4.1", "4.1-4.9",
+     ">4.9")), C = c(21, 17, 11, 1, 4, 25, 35, 5, 2, 6, 36, 14,
+     0, 0, 6, 12))
> tb.c.ex18$symm <- paste(pmin(as.numeric(tb.c.ex18$begin),
+                               as.numeric(tb.c.ex18$end)),
+     pmax(as.numeric(tb.c.ex18$begin), as.numeric(tb.c.ex18$end)),
+     sep = ",")
> tb.t.ex18$symm <- paste(pmin(as.numeric(tb.c.ex18$begin),
+                               as.numeric(tb.c.ex18$end)),
+     pmax(as.numeric(tb.c.ex18$begin), as.numeric(tb.c.ex18$end)),
+     sep = ",")
> tb.c.ex18$symm <- factor(tb.c.ex18$symm, levels = rev(unique(tb.c.ex18$symm)))
> tb.t.ex18$symm <- factor(tb.t.ex18$symm, levels = rev(unique(tb.t.ex18$symm)))
> tb.c.ex18$begin.f <- factor(tb.c.ex18$begin, levels = rev(unique(tb.c.ex18$begin)))
> tb.t.ex18$begin.f <- factor(tb.t.ex18$begin, levels = rev(unique(tb.t.ex18$begin)))
> tb.c.ex18$scores <- as.numeric(tb.c.ex18$end)
> tb.t.ex18$scores <- as.numeric(tb.t.ex18$end)
```

Then we can fit the models and compare how well they fit through their deviances and residuals.

```
> ## Control group
> fit.c.qsymm.ex18 <- glm(C ~ symm + begin.f, family = poisson(log),
+     data = tb.c.ex18)
> fit.c.oqsymm.ex18 <- glm(C ~ symm + scores, family = poisson(log),
+     data = tb.c.ex18)
> ## Treatment group
> fit.c.qsymm.ex18 <- glm(C ~ symm + begin.f, family = poisson(log),
+     data = tb.c.ex18)
> fit.c.oqsymm.ex18 <- glm(C ~ symm + scores, family = poisson(log),
+     data = tb.c.ex18)
```

Control Group

Quasi-symmetry model

	<3.4	3.4–4.1	4.1–4.9	>4.9
<3.4	-Inf	0.00	0.00	0.00
3.4–4.1	0.00	-Inf	-0.70	0.70
4.1–4.9	0.00	0.70	-Inf	-0.70
>4.9	0.00	-0.70	0.70	-Inf

Ordinal quasi-symmetry model

	<3.4	3.4–4.1	4.1–4.9	>4.9
<3.4	Inf	-0.80	0.00	0.00
3.4–4.1	0.80	-Inf	1.00	0.90
4.1–4.9	0.00	-1.00	-Inf	-0.90
>4.9	0.00	-0.90	0.90	-Inf

Treatment Group

Quasi-symmetry model

	<3.4	3.4–4.1	4.1–4.9	>4.9
<3.4	Inf	-1.50	1.60	-0.20
3.4–4.1	1.50	Inf	-0.90	-0.70
4.1–4.9	-1.60	0.90	Inf	0.70
>4.9	0.20	0.70	-0.70	0.00

Ordinal quasi-symmetry model

	<3.4	3.4–4.1	4.1–4.9	>4.9
<3.4	0.00	-0.30	1.30	-0.10
3.4–4.1	0.30	Inf	-1.30	-0.60
4.1–4.9	-1.30	1.30	-Inf	1.00
>4.9	0.10	0.60	-1.00	Inf

Table 5: Residuals for the quasi-symmetry and ordinal quasi-symmetry models applied to the control and treatment groups

Because the residuals for the ordinal quasi-symmetry model are not bigger than 2, and the p-values are both non-significant (0.78 for the control group and 0.56 for the treatment group) we can use this model. It has the advantage that the interpretation is easier: the probability that the second observation is x categories higher than the first observation (*i.e.* that the level of LDL cholesterol increased) is $exp(\beta x)$ times the probability that the first observation is x categories higher than the second observation. For example for the control group the odds ratio that the cholesterol level increased from the 3.4–4.1 to 4.1–4.9 is $exp(\beta 1) = exp(-0.3893) = 0.68$ times the odds that the cholesterol level has reduced. Since $\beta < 0$ we can be confident that the initial cholesterol values are higher than the final ones. This difference is more dramatic for the treatment group; for the same odds ratio the value is only $exp(-1.3059) = 0.27$.

Problem 19

For this problem we can use the data of problem 14, removing the columns that aren't necessary. Then we need to add column for each factor which will be used for the quasi-independence model.

```
> tb.ex19 <- tb.ex14[, 1:3]
> tb.ex19$D1 <- as.numeric(as.numeric(tb.ex19$Resid.1) == 1 & as.numeric(tb.ex19$Resid.2) ==
+     1)
> tb.ex19$D2 <- as.numeric(as.numeric(tb.ex19$Resid.1) == 2 & as.numeric(tb.ex19$Resid.2) ==
+     2)
> tb.ex19$D3 <- as.numeric(as.numeric(tb.ex19$Resid.1) == 3 & as.numeric(tb.ex19$Resid.2) ==
+     3)
> tb.ex19$D4 <- as.numeric(as.numeric(tb.ex19$Resid.1) == 4 & as.numeric(tb.ex19$Resid.2) ==
+     4)
```

As stated in the enunciate, the residuals in the main diagonal for the independence model are around 40, while those for the quasi-Independence model are infinite. This is because in the formula for the standardized residuals (see Agresti 2007, p. 148):

$$\frac{y_i - n_i\hat{\pi}_i}{SE} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{[n_i\hat{\pi}_i(1-\hat{\pi}_i)(1-h_i)]}}$$

the denominator is zero. This is because the quasi-independence model allows agreement (same classification for both observations) beyond independence, thus, in this case yielding a perfect fit for the main diagonal.

Problem 20

First we construct the data:

```
> tb.ex20 <- data.frame(expand.grid(Neu.A = 1:4, Neu.B = 1:4),
+     C = c(38, 33, 10, 3, 5, 11, 14, 7, 0, 3, 5, 3, 1, 0, 6, 10))
> tb.ex20$Neu.A <- factor(tb.ex20$Neu.A, levels = 1:4)
> tb.ex20$Neu.B <- factor(tb.ex20$Neu.B, levels = 1:4)
> tb.ex20$D1 <- as.numeric(tb.ex20$Neu.A == 1 & tb.ex20$Neu.B ==
+     1)
> tb.ex20$D2 <- as.numeric(tb.ex20$Neu.A == 2 & tb.ex20$Neu.B ==
+     2)
> tb.ex20$D3 <- as.numeric(tb.ex20$Neu.A == 3 & tb.ex20$Neu.B ==
+     3)
> tb.ex20$D4 <- as.numeric(tb.ex20$Neu.A == 4 & tb.ex20$Neu.B ==
+     4)
> tb.ex20$symm <- paste(pmin(as.numeric(tb.ex20$Neu.A),
+                       as.numeric(tb.ex20$Neu.B)),
+                  pmax(as.numeric(tb.ex20$Neu.A),
+                       as.numeric(tb.ex20$Neu.B)),
+                  sep = ",")
> tb.ex20$symm <- factor(tb.ex20$symm, levels = rev(unique(tb.ex20$symm)))
```

a) Independence model and standardized residuals:

```
> std.residuals <- function(x){resid(x,type="pearson")/sqrt(1-lm.influence(x)$hat)}
> fit.i.ex20 <- glm(C ~ Neu.A + Neu.B, data = tb.ex20, family = poisson(log))
> xtabs(round(std.residuals(fit.i.ex20), 1) ~ Neu.A + Neu.B, data = tb.ex20)
```

```
     Neu.B
Neu.A    1    2    3    4
    1  4.8 -2.5 -2.2 -2.3
    2  2.3 -0.3 -0.3 -3.0
    3 -3.8  2.4  1.8  1.2
    4 -4.6  0.7  1.1  5.3
```

Note that more than one residual exceeds the value of 3, particularly for the one corresponding to the category 4. This, together with the deviance's p-value (2.22e-11) indicate that the model fits poorly. Now, regarding a possible pattern of agreement only the first and last element of the main diagonal seem to have more observations that the model of independence predicts.

b) First we fit a quasi-independence model (and standardized residuals), *i.e.* one that allows agreement beyond independence.

```
> fit.qi.ex20 <- glm(C ~ Neu.A + Neu.B + D1 + D2 + D3 + D4, data = tb.ex20,
+     family = poisson(log))
> xtabs(round(std.residuals(fit.qi.ex20), 1) ~ Neu.A + Neu.B, data = tb.ex20)
```

```
     Neu.B
Neu.A    1    2    3    4
    1 -Inf  0.3 -0.9  0.5
    2  3.9 -Inf -1.4 -3.6
    3 -1.7 -0.5  Inf  3.5
    4 -1.7  0.4  2.5 -Inf
```

Note that although the fit for the main diagonal is perfect, there are two cells that exhibit lack of fit. The deviance and associated p-value show some improvement but yet a poorly fit (G^2=22.04, p-value=0.0005). As mentioned by Agresti (2007) the quasi-independence model assumes, conditioned on the observer's disagreement, that the X and Y variables are independent, thing does not always occur. Contrary, the quasi-symmetry model assumes other type of association for those cell off the main diagonal, thus we fit that model and its residuals.

```
> fit.qsymm.ex20 <- glm(C ~ symm + Neu.A, data = tb.ex20, family = poisson(log))
> xtabs(round(std.residuals(fit.qsymm.ex20), 1) ~ Neu.A + Neu.B,
+     data = tb.ex20)
```

```
     Neu.B
Neu.A    1    2    3    4
    1  Inf -1.0 -0.5  2.3
    2  1.0 -Inf  0.9 -1.8
    3  0.5 -0.9 -Inf  0.7
    4 -2.3  1.8 -0.7 -Inf
```

Although the residuals show some lack of fit, the model fits the data decently enough (G^2=6,184 on 3 df, p-value = 0.103).

Assuming that the quasi-symmetry model holds, we can calculate the *agreement odds*, for this we need to calculate the fitted counts. Now, for example if we want to calculate the odds that one observer classifies, one subject as 3 rather than 1, when the other also classified the subject as 3 instead of 1 is, 5*38/(9.7*0.3) = 65.2921. Thompson (2007) describes a way of calculating the odds of agreement based on the coefficients of `symm` factor. For the same comparisons this is:

	1	2	3	4
1	38	5	0	1
2	33	11	3	0
3	10	14	5	6
4	3	7	3	10

	1	2	3	4
1	38.0	5.6	0.3	0.2
2	32.4	11.0	2.2	1.3
3	9.7	14.8	5.0	5.5
4	3.8	5.7	3.5	10.0

Table 6: Observed and fitted counts (quasi-symmetry) for problem 20, diagnosis of multiple sclerosis by two neurologists

```
> exp(fit.qsymm.ex20$coefficients["symm3,3"] +
+     fit.qsymm.ex20$coefficients["symm1,1"] -
+     2 * fit.qsymm.ex20$coefficients["symm1,3"])
```

```
 symm3,3
76.44598
```

Note that the values are different, but this is due to round error; if the table of fitted values wasn't rounded we would get the same values as the one obtained by the approach described in Thompson (2007).

c)Cohen's kappa is calculated as follows:

```
> array.ex20 <- xtabs(C ~ Neu.A + Neu.B, data = tb.ex20)
> Kappa(array.ex20, weights = "Fleiss-Cohen")
```

```
            value     ASE     z  Pr(>|z|)
Unweighted 0.2079 0.05046 4.121 3.767e-05
Weighted   0.5246 0.06006 8.735 2.438e-18
```

The interpretation is that, the difference between observed agreement and that expected under independence is more than 50% of the maximum value. We use the "weighted" value because it takes into account the ordinal categories (see Agresti 2007, section 8.5.5).

Note about Kappa It's important to note that nither Agresti (2007) nor Thompson (2007) gives a threshold for the value of kappa.

Problem 21

Data construction,

```
> tb.ex21 <- data.frame(expand.grid(Buy.1 = c("H.Point", "Taster",
+     "Sanka", "Nescafe", "Brim"), Buy.2 = c("H.Point", "Taster",
+     "Sanka", "Nescafe", "Brim")), C = c(93, 9, 17, 6, 10, 17,
+     46, 11, 4, 4, 44, 11, 155, 9, 12, 7, 0, 9, 15, 2, 10, 9,
+     12, 2, 27))
> tb.ex21$D1 <- as.numeric(as.numeric(tb.ex21$Buy.1) == 1 & as.numeric(tb.ex21$Buy.2) ==
+     1)
> tb.ex21$D2 <- as.numeric(as.numeric(tb.ex21$Buy.1) == 2 & as.numeric(tb.ex21$Buy.2) ==
+     2)
> tb.ex21$D3 <- as.numeric(as.numeric(tb.ex21$Buy.1) == 3 & as.numeric(tb.ex21$Buy.2) ==
+     3)
> tb.ex21$D4 <- as.numeric(as.numeric(tb.ex21$Buy.1) == 4 & as.numeric(tb.ex21$Buy.2) ==
+     4)
> tb.ex21$D5 <- as.numeric(as.numeric(tb.ex21$Buy.1) == 5 & as.numeric(tb.ex21$Buy.2) ==
+     5)
```

Then we fit the model and calculate the fitted values

```
> fit.qi.ex21 <- glm(C ~ Buy.1 + Buy.2 + D1 + D2 + D3 + D4 + D5,
+     family = poisson(log), data = tb.ex21)
> xtabs(fitted(fit.qi.ex21) ~ Buy.1 + Buy.2, data = tb.ex21)
```

```
         Buy.2
Buy.1        H.Point     Taster      Sanka    Nescafe       Brim
  H.Point  93.000000  15.567112  40.836359   7.426762  14.169766
  Taster    8.427931  46.000000  13.455863   2.447170   4.669036
  Sanka    19.975643  12.157728 155.000000   5.800212  11.066417
  Nescafe   5.586291   3.399971   8.918958  15.000000   3.094780
  Brim      8.010135   4.875189  12.788819   2.325857  27.000000
```

Finally to calculate the odds ratio we can index the rows and columns that indicates the statement.

```
> oddsratio(xtabs(fitted(fit.qi.ex21) ~ Buy.1 + Buy.2, data = tb.ex21)[c("H.Point",
+     "Taster"), c("Nescafe", "Brim")], log = FALSE)
```

```
 odds ratios for Buy.1 and Buy.2

[1] 1
```

The value of 1 is due to the assumption of the quasi-independence model, that all the values off the main diagonal are independent of each other.

Problem 22

Thompson (2007) describes a method to calculate the Bradley–Terry model, however its cumbersome. Fortunately there is one package called, BradleyTerry2 that simplyfies the calculation. Using this package the commands are:

```
> tb.ex22 <- data.frame(expand.grid(Win = c("Coke", "Pepsi", "C.Coke"),
+     Lose = c("Coke", "Pepsi", "C.Coke")), C = c(0, 29, 31, 20,
+     0, 28, 19, 19, 0))
> library(BradleyTerry2)
> tb2.ex22 <- countsToBinomial(xtabs(C ~ Win + Lose, data = tb.ex22))
> bt.ex22 <- BTm(cbind(win1, win2), player1, player2, ~drink, id = "drink",
+     data = tb2.ex22, )
```

There are a number of things to note. First, the command countsToBinomial converts a matrix to a data frame of "wins" and "loses" for all the pairs. Second, the BTm has a different syntax, first we need to pair the victories and loses through a cbind command (just like in logistic regression), then we need to specify the columns of explanatory variables and then specify an arbitrary name to put as a prefix for the coefficients. This name must match to one for the id argument.
The results of the model are:

```
Bradley Terry model fit by glm.fit

Call:  BTm(outcome = cbind(win1, win2), player1 = player1, player2 = player2,
    formula = ~drink, id = "drink", data = tb2.ex22)

Coefficients:
 drinkPepsi  drinkC.Coke
     0.2837       0.5796

Degrees of Freedom: 3 Total (i.e. Null);  1 Residual
Null Deviance:       6.305
Residual Deviance: 0.2827        AIC: 17.19
```

Note that, as usual, the first category (in this case Coke) was set to zero, thus the ranking is: 1) Clasic Coke, 2) Pepsi and 3) Coke.

b) Now, the probability that between Pepsi and Coke, the latter is choosen is:

```
> exp(-0.2837)/(1 + exp(-0.2837))
```

```
[1] 0.4295469
```

This value is close to the sample proportion $(20/49 = 0.41)$.

Problem 23

Just like the previous problem first we construct the data,

```
> tb.ex23 <- data.frame(expand.grid(Cit = c("Biometrika", "C.Statist.",
+     "JASA", "JRSS.B"), Ctd = c("Biometrika", "C.Statist.", "JASA",
+     "JRSS.B")), C = c(714, 730, 498, 221, 33, 425, 68, 17, 320,
+     813, 1072, 142, 284, 276, 325, 188))
> tb2.ex23 <- countsToBinomial(xtabs(C ~ Ctd + Cit, data = tb.ex23))
```

a) Then we fit the model

```
> (bt.ex23 <- BTm(cbind(win1, win2), player1, player2, ~journal,
+     id = "journal", data = tb2.ex23))
```

```
Bradley Terry model fit by glm.fit

Call:  BTm(outcome = cbind(win1, win2), player1 = player1, player2 = player2,
    formula = ~journal, id = "journal", data = tb2.ex23)

Coefficients:
journalC.Statist.        journalJASA      journalJRSS.B
          -2.9491            -0.4796             0.2690

Degrees of Freedom: 6 Total (i.e. Null);  3 Residual
Null Deviance:          1925
Residual Deviance: 4.293        AIC: 46.39
```

Based on the coefficients the ranking of journals is: 1)JRSS-B, 2) Biometrika, 3) JASA and 4)C. Statist.

b) The probability that between C. Statist and JRSS-B the latter is choosen is:

```
> exp(0.269 + 2.9491)/(1 + exp(0.269 + 2.9491))
```

```
[1] 0.9615098
```

Problem 24

Data input

```
> tb.ex24 <- data.frame(expand.grid(Win = c("Clijsters", "Davenport",
+     "Pierce", "S.Williams", "V.Williams"), Lose = c("Clijsters",
+     "Davenport", "Pierce", "S.Williams", "V.Williams")), C = c(0,
+     2, 1, 2, 3, 6, 0, 2, 2, 2, 3, 0, 0, 2, 2, 0, 2, 0, 0, 2,
+     2, 4, 1, 2, 0))
> library(BradleyTerry2)
> tb2.ex24 <- countsToBinomial(xtabs(C ~ Win + Lose, data = tb.ex24))
```

Model fit

```
> (bt.ex24 <- BTm(cbind(win1, win2), player1, player2, ~player,
+     id = "player", data = tb2.ex24))
```

```
Bradley Terry model fit by glm.fit

Call:  BTm(outcome = cbind(win1, win2), player1 = player1, player2 = player2,
    formula = ~player, id = "player", data = tb2.ex24)
```

```
Coefficients:
 playerDavenport      playerPierce  playerS.Williams
         -0.4470           -0.6249            0.3918
playerV.Williams
         -0.1674

Degrees of Freedom: 10 Total (i.e. Null);  6 Residual
Null Deviance:          12.68
Residual Deviance: 10.12         AIC: 32.07
```

a) Based on the model, the ranking of players is: 1) S. Williams, 2) Clijsters, 3) V. Williams, 4) Davenport and 5) Pierce.

b) The probability that S. Williams beats her sister V. Williams is exp(0.3918-(-0.1674))/(1+exp(0.3918-(-0.1674))) = 0.64 or 64%. The sample proportion is 2/4 = 0.5 or 50%. Note that the model probability is higher than the sample proportion, in other words S. Williams has greater probability of winning.

c) *I don't know how to find the SE of the difference*

d) The results from the model give the null and residual deviances and their degrees of freedom so one just have to calculate the binomial probability of the difference of deviances and degrees of freedom. Alternatively one can use the function `anova` with the fit-variable (in this case `bt.ex24`)

```
> anova(bt.ex24, test = "Chisq")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(win1, win2)

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                      10     12.678
player  4   2.5611         6     10.117   0.6337
```

Based on the p-value we can not reject the null hypothesis that the simpler model holds, or in other words that there are no differences among the players (marginal homogeneity).

Problem 25

Data construction

```
> tb8.9 <- data.frame(expand.grid(Win = c("Agassi", "Federer",
+     "Henman", "Hewitt", "Roddick"), Lose = c("Agassi", "Federer",
+     "Henman", "Hewitt", "Roddick")), C = c(0, 6, 0, 0, 0, 0,
+     0, 1, 0, 0, 0, 3, 0, 2, 1, 1, 9, 0, 0, 2, 1, 5, 1, 3, 0))
> tb28.9 <- countsToBinomial(xtabs(C ~ Win + Lose, data = tb8.9))
```

Model fit

```
> (bt.89 <- BTm(cbind(win1, win2), player1, player2, ~player, id = "player",
+     data = tb28.9))
```

```
Bradley Terry model fit by glm.fit

Call:  BTm(outcome = cbind(win1, win2), player1 = player1, player2 = player2,
    formula = ~player, id = "player", data = tb28.9)

Coefficients:
playerFederer   playerHenman   playerHewitt  playerRoddick
       2.4327        -1.2613        -0.8755        -1.4489

Degrees of Freedom: 9 Total (i.e. Null);  5 Residual
Null Deviance:	    34.52
Residual Deviance: 8.191 	AIC: 21.43
```

a) Probability of a victory of Agassi over Henman is exp(0-(-1.2613))/(1+exp(0-(-1.2613))) = 0.78.

b) The likelihood statistic for the null hypothesis that $H_0 : \beta_1 = \ldots = \beta_5$ is

```
> anova(bt.89, test = "Chisq")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(win1, win2)

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       9     34.519
player  4   26.328         5      8.191 2.717e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There is strong evidence to reject the null hypothesis and say that there are differences among the players.

Problem 26

Agresti (2007) defines the probability that i is preferred over j as:

$$\Pi_{ij} = \frac{exp(\hat{\beta}_i - \hat{\beta}_j)}{1 + exp(\hat{\beta}_i - \hat{\beta}_j)}$$

In order to Π be greater than 0.5, the numerator must be positive which implies that $\hat{\beta}_i > \hat{\beta}_j$. Now, if A is preferred over B and B is preferred over C then $\beta_a > \beta_b$ and $\beta_b > \beta_c$. Therefore it's impossible that

$\beta_c > \beta_a$.

Problem 27

a) this is explained in Thompson (2007, p. 198 and in the notes) with the following equation

$$\begin{aligned} log\frac{\mu_{ij}}{\mu_{ba}} &= (\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}) - (\lambda + \lambda_j^X + \lambda_i^Y + \lambda_{ij}) \\ &= (\lambda_i^X - \lambda_i^Y) - (\lambda_j^X - \lambda_j^Y) = \beta_i - \beta_j \end{aligned}$$

b) Now, if $\lambda_i^X = \lambda_i^Y$, the terms in the above equation cancel out, *i.e.* the equation reduces to

$$log\frac{\mu_{ij}}{\mu_{ba}} = 0$$

which is the equation for the independence model.

c) If λ_{ij} is zero when $i \neq j$ then we would add a term for each time $i = j$ (main diagonal), term that is represented by $\sigma_i I(i = j)$ in the equation of quasi-independence:

$$log\mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \sigma_i I(i = j)$$

Problem 28

In Agresti (2002, Section 10.2.6 p. 419) it's explained how to construct the data frame to fit the logistic model for matched pairs. There it's said that Y^* will always be 1, while the values for X^* will be 1, for cell n_{21}, -1 for cell n_{12} and 0 for the others. So with this in mind we contruct the data frame.

```
> (tb8.1 <- matrix(c(227, 107, 132, 678), ncol = 2, dimnames = list(Taxes = c("yes",
+     "no"), Std = c("yes", "no"))))
```

```
     Std
Taxes yes  no
  yes 227 132
  no  107 678
```

```
> tb.ex28 <- data.frame(Y = c(1, 1, 1), X = c(1, -1, 0))
> tb.ex28a <- rbind(tb.ex28[rep(1, 107), ], tb.ex28[rep(2, 132),
+     ], tb.ex28[rep(3, sum(diag(tb8.1))), ])
> tb.ex28b <- rbind(tb.ex28[rep(1, 107), ], tb.ex28[rep(2, 132),
+     ], tb.ex28[rep(3, 2), ])
```

Then we can fit the ordinal logistic model with no intercept (-1)

```
> summary(fit.ex28 <- glm(Y ~ X - 1, family = binomial, data = tb.ex28a))
```

```
Call:
glm(formula = Y ~ X - 1, family = binomial, data = tb.ex28a)

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
```

```
X  -0.2100     0.1301  -1.614    0.106

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1585.9  on 1144  degrees of freedom
Residual deviance: 1583.3  on 1143  degrees of freedom
AIC: 1585.3

Number of Fisher Scoring iterations: 3
```

a) As it's shown the estimate value is -0.21 with a standard error of 0.13. Note that the β estimate represents the log odds ratio of being in favor of a cut in living standards instead of a raise in taxes. If we want the opposite we need to change the sign of the estimate (see note 10)

b)Agresti (2007) mentions that the values for the cells n_{11} and n_{22} do not contribute to the estimate. To test this we can modify the number of these cells as it was done for `tb.ex28b` in the above commands. Then we can fit the model with these new data and compare the estimate.

```
> summary(glm(Y ~ X - 1, family = binomial, data = tb.ex28b))
```

```
Call:
glm(formula = Y ~ X - 1, family = binomial, data = tb.ex28b)

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
X  -0.2100     0.1301  -1.614    0.106

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 334.10  on 241  degrees of freedom
Residual deviance: 331.48  on 240  degrees of freedom
AIC: 333.48

Number of Fisher Scoring iterations: 3
```

Note that the value of the estimate as well as its standard error do not change. Also note that these cells are not taken into account for the McNemar test.

References

Agresti, Alan (2002). *Categorical Data Analysis*. 2nd ed. New York: Wiley-Interscience. ISBN: 978-0-471-36093-3.

— (2007). *An Introduction to Categorical Data Analysis*. 2nd ed. Hoboken NJ: Wiley-Interscience. ISBN: 978-0-471-22618-5.

Thompson, Laura A. (2007). *S-plus (and R) Manual to Accompany Agresti's "Categorical Data Analysis" (2002)*.