

Epidemiologic Analysis: A Case-Oriented Approach

R code from Tomás Aragón, modified 2005-01-27

Chapter 2

Approaches to assessing measures of association (odds ratio and risk ratio):

1. Estimate a common measure of association (MA)?

1.1. Woolf estimate

1.2. Mantel-Haenszel estimate

1.3. Maximum likelihood estimate

2. Do stratum-specific MAs differ from the common MA?

2.1. Woolf's test for homogeneity

2.2. Regression model approach (Logistic and Poisson)

a) No interaction vs. interaction model

b) Goodness of fit approach

3. Does the common MA differ from 1 (Test for association)?

3.1. Standard normal z test

3.2. Regression coefficient

4. Confounding

R code running

```
> ##Chapter2: p. 13
>
> ##PART I: ODDS RATIO,
>
> ##lowbwt.2.data
> lbwdat <- read.table("http://www.medepi.net/selvin/lowbwt.2.data", header =
FALSE, sep="")
> names(lbwdat) <- c("bwt", "smk", "eth", "count")
> attr(lbwdat, "dict") <- c("bwt = birth weight (1: <2500 g; 0: >=2500 g)",
+ "smk = smoking exposure (1: Yes; 0: otherwise)",
+ "eth = race/ethnicity (1: white; 2: African American;
3: Hispanic; 4: Asian)",
+ "count = cell frequency")
> str(lbwdat)
`data.frame': 16 obs. of 4 variables:
 $ bwt : int 0 1 0 1 0 1 0 1 0 1 ...
 $ smk : int 0 0 1 1 0 0 1 1 0 0 ...
 $ eth : int 1 1 1 1 2 2 2 2 3 3 ...
 $ count: int 3520 169 832 98 686 55 227 54 926 61 ...
- attr(*, "dict")= chr "bwt = birth weight (1: <2500 g; 0: >=2500 g)" "smk =
```

```

1 smoking exposure (1: Yes; 0: otherwise)" "eth = race/ethnicity (1: white; 2:
2 African American; 3: Hispanic; 4: Asian)" "count = cell frequency"
3 > lbwdat
4     bwt smk eth count
5 1     0  0  1 3520
6 2     1  0  1  169
7 3     0  1  1  832
8 4     1  1  1   98
9 5     0  0  2  686
10 6     1  0  2   55
11 7     0  1  2  227
12 8     1  1  2   54
13 9     0  0  3  926
14 10    1  0  3   61
15 11    0  1  3   85
16 12    1  1  3   11
17 13    0  0  4 1936
18 14    1  0  4   90
19 15    0  1  4  102
20 16    1  1  4    7
21 >
22 > ##re-create original data set on p. 14
23 > bwt <- rep(lbwdat$bwt, lbwdat$count)
24 > smk <- rep(lbwdat$smk, lbwdat$count)
25 > eth <- rep(lbwdat$eth, lbwdat$count)
26 >
27 > bwt <- factor(bwt, levels = 1:0, labels = c("<2500g", ">=2500g"))
28 > smk <- factor(smk, levels = 1:0, labels = c("Smokers", "Nonsmokers"))
29 > eth <- factor(eth, levels = 1:4, labels = c("white", "African American",
30 "Hispanic", "Asian"))
31 >
32 > lbw <- data.frame(bwt, smk, eth)
33 >
34 > ##2x2 table
35 > table(Smoking = lbw$smk, "Birth weight" = lbw$bwt)
36           Birth weight
37 Smoking    <2500g >=2500g
38   Smokers      170   1246
39  Nonsmokers   375   7068
40 >
41 > ##Table 2.1 p. 16
42 > ##2x2x4 table: 3-dimensional array (best for manipulation)
43 > tab2.1 <- table(Smoking = lbw$smk, "Birth weight" = lbw$bwt, Ethnicity = lbw$eth)
44 > tab2.1
45 , , Ethnicity = white
46
47           Birth weight
48 Smoking    <2500g >=2500g
49   Smokers      98   832
50  Nonsmokers  169  3520
51
52 , , Ethnicity = African American
53
54           Birth weight

```

```

1 Smoking      <2500g >=2500g
2   Smokers      54    227
3   Nonsmokers   55    686
4
5 , , Ethnicity = Hispanic
6
7           Birth weight
8 Smoking      <2500g >=2500g
9   Smokers      11    85
10  Nonsmokers   61   926
11
12 , , Ethnicity = Asian
13
14           Birth weight
15 Smoking      <2500g >=2500g
16  Smokers       7   102
17  Nonsmokers   90  1936
18
19 >
20 > ##alternative view of arrays: flat contingency table (easier to view)
21 > tab2.1b <- ftable(lbw$eth, lbw$smk, lbw$bwt)
22 > tab2.1b
23
24           <2500g >=2500g
25 white           Smokers      98    832
26                 Nonsmokers  169   3520
27 African American Smokers      54    227
28                 Nonsmokers  55    686
29 Hispanic        Smokers      11    85
30                 Nonsmokers  61   926
31 Asian           Smokers       7   102
32                 Nonsmokers  90  1936
33 >

```

$$\hat{or}_i = \frac{a_i d_i}{b_i c_i}$$

$$S^2_{\log(\hat{or}_i)} = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

```

36 > ##calculate strata-specific odds ratio and CIs for Table 2.2
37 > ai <- tab2.1[1,1,]
38 > bi <- tab2.1[1,2,]
39 > ci <- tab2.1[2,1,]
40 > di <- tab2.1[2,2,]
41 > ni <- ai + bi + ci + di
42 > ori <- (ai*di)/(bi*ci)
43 > conf.level <- 0.95
44 > Z <- qnorm(0.5*(1 + conf.level))
45 > log.ori <- log(ori)
46 > var.log.ori <- 1/ai + 1/bi + 1/ci + 1/di
47 > sd.log.ori <- sqrt(var.log.ori)
48 > lower <- exp(log.ori - Z*sd.log.ori)
49 > upper <- exp(log.ori + Z*sd.log.ori)

```

```

1 >
2 >
3 > ##Table 2.2 p. 16
4 > tab2.2 <- round(cbind(ni, ori, lower, upper), 3)
5 > tab2.2
6           ni  ori lower upper
7 White      4619 2.453 1.892 3.182
8 African American 1022 2.967 1.980 4.446
9 Hispanic    1083 1.965 0.996 3.875
10 Asian      2135 1.476 0.667 3.267
11 >
12 > wi <- 1/var.log.ori ##weights are inverse of variance
13 >
14 > ##Table 2.3 p. 16
15 > tab2.3 <- round(cbind(ni, log.ori, var.log.ori, wi), 3)
16 > tab2.3
17           ni log.ori var.log.ori  wi
18 White      4619  0.897  0.018 56.795
19 African American 1022  1.088  0.043 23.494
20 Hispanic    1083  0.675  0.120  8.323
21 Asian      2135  0.390  0.164  6.087
22 >
23 > ##First, do the odds ratios systematically differ among the ethnic groups?
24 > ##More specifically, we will look at the log-odds ratios
25 > ##Let's evaluate the amount of variability among the observed odds ratios
26 > ##relative to the mean log-odds ratio
27 >
28

$$\overline{\log(or)} = \frac{\sum w_i \log(\hat{or}_i)}{\sum w_i}$$

29 > log.or.bar <- sum(wi * log.ori)/sum(wi)
30 > log.or.bar
31 [1] 0.8924434
32 >
33

$$z_i = \frac{\log(\hat{or}_i) - \overline{\log(or)}}{S_{\log(\hat{or}_i)}}$$

34 > zi <- (log.ori - log.or.bar)/sd.log.ori
35 > zi
36           white African American           Hispanic           Asian
37           0.0377501           0.9458393           -0.6266156           -1.2408215
38 >
39 > ##[Method 1] woolf's chi square test for homogeneity
40 >
41

$$\chi^2 = \sum z_i^2$$

42 > chi2.w <- sum(zi^2) ##with k-1 d.f.
43 > chi2.w
44 [1] 2.828322
45 >
46 > sum(wi*(log.ori - log.or.bar)^2)
47 [1] 2.828322
48 >

```

```

1 > 1 - pchisq(q = chi2.w, df = length(zi) - 1)
2 [1] 0.4188588
3 > ##equivalent and preferred
4 > pchisq(q = chi2.w, df = length(zi) - 1, lower.tail = FALSE)
5 [1] 0.4188588
6 >
7 > ##no evidence that odds ratios differ systematically among ethnic groups
8 > ##therefore, calculate common odds ratios
9 >
10 > ##woolf's estimate
11 > or.w <- exp(log.or.bar)
12 > or.w
13 [1] 2.441087
14 > ##now test whether or.w differs from 1 (or log.or.bar differs from 0)
15 >
16

$$S_{\log(or)}^2 = \frac{1}{\sum w_i}$$

17 > var.log.or.bar <- 1/sum(wi)
18 >
19

$$z = \frac{\overline{\log(or)} - \log(1)}{S_{\log(or)}}$$

20 > z <- (log.or.bar - log(1))/sqrt(var.log.or.bar)
21 > z
22 [1] 8.684678
23 > ##P(Z>=z)
24 > pnorm(z, lower.tail=FALSE)
25 [1] 1.899089e-18
26 >
27 > ##CI for common odds ratio
28 > conf.level <- 0.95
29 > Z <- qnorm(0.5 *(1 + conf.level))
30 > lower <- exp(log.or.bar - Z*sqrt(var.log.or.w))
31 > lower
32 [1] 1.995782
33 > upper <- exp(log.or.bar + Z*sqrt(var.log.or.w))
34 > upper
35 [1] 2.985749
36 >
37 > ##Mantel-Haenszel estimate
38 > or.MH <- sum(ai*di/ni)/sum(bi*ci/ni)
39 > or.MH
40 [1] 2.448211
41 >
42 > ##MLE estimate
43 > ##comes from logistic model (below)
44 >
45 >
46 > ##A natural summary, p. 19
47 >
48 > ##Logistic model approach, p. 22
49 >

```

```

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

```

$$p_{ij} = \frac{1}{1 + e^{-\log\text{-odds}_{ij}}}$$

$$\log\text{-odds}_{ij} = a + b x_i + c_1 y_{1j} + c_2 y_{2j} + c_3 y_{3j} \quad i=0,1 \text{ and } j=0,1,2,3$$

```

> ##[Method 2]: Test for homogeneity (interaction)
> ##original data
>
> m1 <- glm(bwt ~ smk + factor(eth), weight = count, data = lbwdat, family =
binomial(link = logit))
> summary(m1)

Call:
glm(formula = bwt ~ smk + factor(eth), family = binomial(link = logit),
    data = lbwdat, weights = count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-18.209  -10.709   0.569   17.114   32.260

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.03193    0.07257  -41.779 < 2e-16 ***
smk           0.88494    0.10109   8.754 < 2e-16 ***
factor(eth)2  0.59819    0.12058   4.961 7.01e-07 ***
factor(eth)3  0.27922    0.13908   2.008  0.0447 *
factor(eth)4 -0.07952    0.12438  -0.639  0.5226
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4095.1 on 15 degrees of freedom
Residual deviance: 3983.1 on 11 degrees of freedom
AIC: 3993.1

Number of Fisher Scoring iterations: 6
>
> m2 <- glm(bwt ~ smk+factor(eth)+smk*factor(eth), weight = count, data = lbwdat,
family = binomial(link = logit))
> summary(m2)

Call:
glm(formula = bwt ~ smk + factor(eth) + smk * factor(eth), family = binomial(link =
logit),
    data = lbwdat, weights = count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-18.17  -10.43   1.26   17.29   32.28

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

```

```

1 (Intercept)      -3.03632      0.07875 -38.558 < 2e-16 ***
2 smk              0.89745      0.13269   6.763 1.35e-11 ***
3 factor(eth)2     0.51277      0.16075   3.190 0.00142 **
4 factor(eth)3     0.31632      0.15387   2.056 0.03980 *
5 factor(eth)4    -0.03225      0.13352  -0.242 0.80913
6 smk:factor(eth)2 0.19013      0.24530   0.775 0.43829
7 smk:factor(eth)3 -0.22221     0.37115  -0.599 0.54937
8 smk:factor(eth)4 -0.50795     0.42649  -1.191 0.23366
9 ---
10 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
11
12 (Dispersion parameter for binomial family taken to be 1)
13
14 Null deviance: 4095.1 on 15 degrees of freedom
15 Residual deviance: 3980.1 on 8 degrees of freedom
16 AIC: 3996.1
17
18 Number of Fisher Scoring iterations: 6
19
20 >
21 > ##list glm model values
22 > names(m1)
23 [1] "coefficients"      "residuals"      "fitted.values"
24 [4] "effects"           "R"               "rank"
25 [7] "qr"                "family"          "linear.predictors"
26 [10] "deviance"          "aic"             "null.deviance"
27 [13] "iter"              "weights"         "prior.weights"
28 [16] "df.residual"       "df.null"         "y"
29 [19] "converged"         "boundary"        "model"
30 [22] "call"              "formula"         "terms"
31 [25] "data"              "offset"          "control"
32 [28] "method"            "contrasts"      "xlevels"
33 >
34 > ##Test for homogeneity (interaction)
35 > m12.resid.dev <- m1$deviance - m2$deviance
36 > m12.resid.dev
37 [1] 3.04175
38 > m12.df.resid <- m1$df.residual - m2$df.residual
39 > m12.df.resid
40 [1] 3
41 > pchisq(q = m12.resid.dev, df = m12.df.resid, lower.tail = FALSE)
42 [1] 0.3852328
43 > ##short cut to compare nested models
44 > anova(m1, m2, test = "Chisq")
45 Analysis of Deviance Table
46
47 Model 1: bwt ~ smk + factor(eth)
48 Model 2: bwt ~ smk + factor(eth) + smk * factor(eth)
49 Resid. Df Resid. Dev Df Deviance P(>|Chi|)
50 1 11 3983.1
51 2 8 3980.1 3 3.0 0.4
52 >
53 >
54 > ##expanded full data

```

```

1 > m1full <- glm(bwt~smk + eth, data = lbw, family = binomial(link = logit))
2 > summary(m1full)
3
4 Call:
5 glm(formula = bwt ~ smk + eth, family = binomial(link = logit),
6     data = lbw)
7
8 Deviance Residuals:
9     Min       1Q   Median       3Q      Max
10  -2.5120   0.2952   0.3069   0.3516   0.6208
11
12 Coefficients:
13             Estimate Std. Error z value Pr(>|z|)
14 (Intercept)    2.14698    0.09130  23.516 < 2e-16 ***
15 smkNonsmokers    0.88494    0.10109   8.754 < 2e-16 ***
16 ethAfrican American -0.59819    0.12058  -4.961 7.01e-07 ***
17 ethHispanic    -0.27922    0.13908  -2.008  0.0447 *
18 ethAsian        0.07952    0.12437   0.639  0.5226
19 ---
20 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
21
22 (Dispersion parameter for binomial family taken to be 1)
23
24 Null deviance: 4095.1 on 8858 degrees of freedom
25 Residual deviance: 3983.1 on 8854 degrees of freedom
26 AIC: 3993.1
27
28 Number of Fisher Scoring iterations: 5
29
30 >
31 > m2full <- glm(bwt~smk + eth + smk*eth, data = lbw, family = binomial(link =
32 logit))
33 > summary(m2full)
34
35 Call:
36 glm(formula = bwt ~ smk + eth + smk * eth, family = binomial(link = logit),
37     data = lbw)
38
39 Deviance Residuals:
40     Min       1Q   Median       3Q      Max
41  -2.4956   0.3015   0.3062   0.3572   0.6533
42
43 Coefficients:
44             Estimate Std. Error z value Pr(>|z|)
45 (Intercept)    2.13886    0.10680  20.027 < 2e-16 ***
46 smkNonsmokers    0.89745    0.13269   6.764 1.35e-11 ***
47 ethAfrican American -0.70290    0.18528  -3.794 0.000148 ***
48 ethHispanic    -0.09411    0.33776  -0.279 0.780529
49 ethAsian        0.54020    0.40505   1.334 0.182316
50 smkNonsmokers:ethAfrican American 0.19013    0.24529   0.775 0.438285
51 smkNonsmokers:ethHispanic -0.22221    0.37115  -0.599 0.549373
52 smkNonsmokers:ethAsian -0.50795    0.42649  -1.191 0.233653
53 ---
54 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

1
2 (Dispersion parameter for binomial family taken to be 1)
3
4 Null deviance: 4095.1 on 8858 degrees of freedom
5 Residual deviance: 3980.1 on 8851 degrees of freedom
6 AIC: 3996.1
7
8 Number of Fisher Scoring iterations: 5
9
10 >
11 > anova(m1full, m2full, test = "Chisq")
12 Analysis of Deviance Table
13
14 Model 1: bwt ~ smk + eth
15 Model 2: bwt ~ smk + eth + smk * eth
16 Resid. Df Resid. Dev Df Deviance P(>|Chi|)
17 1 8854 3983.1
18 2 8851 3980.1 3 3.0 0.4
19 >
20 >
21 > ##[Method 3] Chi-square good-of-fit approach
22 > ##We will use m1 (no interaction model)
23 > ##goal is to replicate table 2.6
24 > ##to get fitted log-odds
25 > m1$linear.predictors
26 1 2 3 4 5 6 7 8
27 -3.031927 -3.031927 -2.146983 -2.146983 -2.433732 -2.433732 -1.548788 -1.548788
28 9 10 11 12 13 14 15 16
29 -2.752702 -2.752702 -1.867759 -1.867759 -3.111445 -3.111445 -2.226501 -2.226501
30 > predict(m1)
31 1 2 3 4 5 6 7 8
32 -3.031927 -3.031927 -2.146983 -2.146983 -2.433732 -2.433732 -1.548788 -1.548788
33 9 10 11 12 13 14 15 16
34 -2.752702 -2.752702 -1.867759 -1.867759 -3.111445 -3.111445 -2.226501 -2.226501
35 >
36 > ##to get fitted probabilities
37 > 1/(1 + exp(-predict(m1)))
38 1 2 3 4 5 6 7
39 0.04600420 0.04600420 0.10461345 0.10461345 0.08063639 0.08063639 0.17526133
40 8 9 10 11 12 13 14
41 0.17526133 0.05993423 0.05993423 0.13380124 0.13380124 0.04263764 0.04263764
42 15 16
43 0.09739578 0.09739578
44 > m1$fitted.values
45 1 2 3 4 5 6 7
46 0.04600420 0.04600420 0.10461345 0.10461345 0.08063639 0.08063639 0.17526133
47 8 9 10 11 12 13 14
48 0.17526133 0.05993423 0.05993423 0.13380124 0.13380124 0.04263764 0.04263764
49 15 16
50 0.09739578 0.09739578
51 > fitted(m1)
52 1 2 3 4 5 6 7
53 0.04600420 0.04600420 0.10461345 0.10461345 0.08063639 0.08063639 0.17526133
54 8 9 10 11 12 13 14

```

```

1 0.17526133 0.05993423 0.05993423 0.13380124 0.13380124 0.04263764 0.04263764
2      15      16
3 0.09739578 0.09739578
4 >
5 >
6 > ##get marginal totals
7 > mtot <- tapply(lbwdat$count, list(lbwdat$smk, lbwdat$eth), sum)
8 > mtot
9      1  2  3  4
10 0 3689 741 987 2026
11 1  930 281  96 109
12 > mtot <- as.vector(mtot)
13 > mtot
14 [1] 3689  930  741  281  987  96 2026 109
15 > mtot <- rep(mtot, rep(2, length(mtot)))
16 > mtot
17 [1] 3689 3689  930  930  741  741  281  281  987  987  96  96 2026 2026 109
18 [16] 109
19 > cellprob <- abs(rep(c(1,0), nrow(lbwdat)/2) - fitted(m1))
20 > cellprob
21      1      2      3      4      5      6      7
22 0.95399580 0.04600420 0.89538655 0.10461345 0.91936361 0.08063639 0.82473867
23      8      9     10     11     12     13     14
24 0.17526133 0.94006577 0.05993423 0.86619876 0.13380124 0.95736236 0.04263764
25     15     16
26 0.90260422 0.09739578
27 > ei <- cellprob*mtot
28 >
29 > tab2.6 <- cbind(lbwdat[,1:3], Total = mtot, pij = cellprob,
30 +               oi = lbwdat[,4], ei = ei)
31 > tab2.6
32      bwt smk eth Total      pij      oi      ei
33 1  0  0  1  3689 0.95399580 3520 3519.29051
34 2  1  0  1  3689 0.04600420 169 169.70949
35 3  0  1  1  930 0.89538655 832 832.70949
36 4  1  1  1  930 0.10461345  98  97.29051
37 5  0  0  2  741 0.91936361 686 681.24843
38 6  1  0  2  741 0.08063639  55  59.75157
39 7  0  1  2  281 0.82473867 227 231.75157
40 8  1  1  2  281 0.17526133  54  49.24843
41 9  0  0  3  987 0.94006577 926 927.84492
42 10 1  0  3  987 0.05993423  61  59.15508
43 11 0  1  3   96 0.86619876  85  83.15508
44 12 1  1  3   96 0.13380124  11  12.84492
45 13 0  0  4 2026 0.95736236 1936 1939.61614
46 14 1  0  4 2026 0.04263764  90  86.38386
47 15 0  1  4  109 0.90260422 102  98.38386
48 16 1  1  4  109 0.09739578  7  10.61614
49 >
50 > ##Pearson chi-square goodness-of-fit statistic, p. 24
51 > chi2 <- sum((tab2.6[, "oi"] - tab2.6[, "ei"])^2/tab2.6[, "ei"])
52 > chi2
53 [1] 2.865653
54 > ##recall that d.f. = (r-1)(c-1)(d-1) for 3-way array

```

```

1 > pchisq(chi2, df = 3, lower.tail = FALSE)
2 [1] 0.412806
3 >
4 > ##assess impact of smoking, p. 24 (bottom)
5 > summary(m1)$coef
6           Estimate Std. Error    z value    Pr(>|z|)
7 (Intercept) -3.03192659  0.0725703 -41.7791643 0.000000e+00
8 smk          0.88494319  0.1010934  8.7537183 2.064361e-18
9 factor(eth)2 0.59819497  0.1205789  4.9610250 7.012218e-07
10 factor(eth)3 0.27922446  0.1390784  2.0076771 4.467762e-02
11 factor(eth)4 -0.07951792  0.1243799 -0.6393149 5.226181e-01
12 > bhat <- summary(m1)$coef["smk", "Estimate"]
13 > bhat
14 [1] 0.8849432
15 > Sbhat <- summary(m1)$coef["smk", "Std. Error"]
16 > Sbhat
17 [1] 0.1010934
18 > z <- bhat/Sbhat
19 > z
20 [1] 8.753718
21 > ##P(Z>=z | b = 0)
22 > pnorm(z, lower.tail = FALSE)
23 [1] 1.032180e-18
24 >
25 > ##MLE
26 > log.or.bar
27 [1] 0.8924434
28 > or.MLE <- exp(log.or.bar)
29 > or.MLE
30 [1] 2.441087
31 >
32 > ##confounding: compare crude and summary odds ratios
33 > or.crude <- (sum(ai)*sum(di))/(sum(bi)*sum(ci))
34 > cbind(or.crude, or.w, or.MH, or.MLE)
35       or.crude  or.w  or.MH  or.MLE
36 [1,] 2.571557 2.441087 2.448211 2.441087
37 >
38 >
39 > ##PART II: RELATIVE RISK
40 > ##chd.2.data
41 > chddat <- read.table("http://www.medepi.net/selvin/chd.2.data", header = FALSE,
42 sep="")
43 > names(chddat) <- c("chd", "beh", "wt", "count")
44 > attr(chddat, "dict") <- c("chd = coronary event: 1 = yes, 0 = no",
45 +                           "beh = A-type or B-type behavior: 1 = A, 0 = B",
46 +                           "wt = body weight category: 1: <150; 2: 150-160lb; 3:
47 160-170lb; 4: 170-180lb; 5: >180lb",
48 +                           "count = cell frequencies")
49 > str(chddat)
50 `data.frame`: 20 obs. of 4 variables:
51 $ chd : int 0 1 0 1 0 1 0 1 0 1 ...
52 $ beh : int 0 0 1 1 0 0 1 1 0 0 ...
53 $ wt : int 1 1 1 1 2 2 2 2 3 3 ...
54 $ count: int 305 10 253 22 270 10 235 21 297 21 ...

```

```

1 - attr(*, "dict")= chr "chd = coronary event: 1 = yes, 0 = no" "beh = A-type or
2 B-type behavior: 1 = A, 0 = B" "wt = body weight category: 1: <150; 2: 150-160lb;
3 3: 160-170lb; 4: 170-180lb; 5: >180lb" "count = cell frequencies"
4 > chddat
5   chd beh wt count
6 1    0  0  1  305
7 2    1  0  1  10
8 3    0  1  1  253
9 4    1  1  1  22
10 5    0  0  2  270
11 6    1  0  2  10
12 7    0  1  2  235
13 8    1  1  2  21
14 9    0  0  3  297
15 10   1  0  3  21
16 11   0  1  3  297
17 12   1  1  3  29
18 13   0  0  4  253
19 14   1  0  4  19
20 15   0  1  4  248
21 16   1  1  4  47
22 17   0  0  5  361
23 18   1  0  5  19
24 19   0  1  5  378
25 20   1  1  5  59
26 >
27 > ##re-create original data set
28 > chd <- rep(chddat$chd, chddat$count)
29 > beh <- rep(chddat$"beh", chddat$count)
30 > wt <- rep(chddat$wt, chddat$count)
31 >
32 > chd <- factor(chd, levels = 1:0, labels = c("CHD", "No CHD"))
33 > beh <- factor(beh, levels = 1:0, labels = c("Type A", "Type B"))
34 > wt <- factor(wt, levels = 1:5, labels = c("<150lb", "150-160lb", "160-170lb",
35 "170-180lb", ">180lb"))
36 >
37 > ##create table 2.7, p. 30
38 > tab2.7 <- table(beh, chd, wt)
39 > tab2.7
40 , , wt = <150lb
41
42       chd
43 beh    CHD No CHD
44 Type A  22 253
45 Type B  10 305
46
47 , , wt = 150-160lb
48
49       chd
50 beh    CHD No CHD
51 Type A  21 235
52 Type B  10 270
53
54 , , wt = 160-170lb

```

```

1
2      chd
3 beh      CHD No CHD
4   Type A  29 297
5   Type B  21 297
6
7 , , wt = 170-180|b
8
9      chd
10 beh     CHD No CHD
11  Type A  47 248
12  Type B  19 253
13
14 , , wt = >180|b
15
16      chd
17 beh     CHD No CHD
18  Type A  59 378
19  Type B  19 361
20
21 >
22 > ##alternative view: flat contingency table
23 > tab2.7b <- ftable(wt, beh, chd)
24 > tab2.7b
25                chd CHD No CHD
26 wt      beh
27 <150|b  Type A      22   253
28         Type B      10   305
29 150-160|b Type A      21   235
30         Type B      10   270
31 160-170|b Type A      29   297
32         Type B      21   297
33 170-180|b Type A      47   248
34         Type B      19   253
35 >180|b  Type A      59   378
36         Type B      19   361
37 >
38 > ##create table 2.8, p. 31
39 > ai <- tab2.7[1,1,]
40 > bi <- tab2.7[1,2,]
41 > ci <- tab2.7[2,1,]
42 > di <- tab2.7[2,2,]
43 > ni <- ai + bi + ci + di
44 > Pi <- ai/(ai+bi) ##Pr(disease | exposed)
45 > pi <- ci/(ci+di) ##Pr(disease | not exposed)
46 > rri <- Pi/pi      ##stratum-specific risk ratio
47 > log.rri <- log(rri)
48 > var.log.rri <- ((1 - Pi)/ai) + ((1 - pi)/ci)
49 > sd.log.rri <- sqrt(var.log.rri)
50 > wi <- 1/var.log.rri
51 >
52 > tab2.8 <- cbind(ni, rri, log.rri, var.log.rri, wi)
53 > round(tab2.8, 3)
54          ni    rri log.rri var.log.rri    wi

```

```

1 <150lb 590 2.520 0.924 0.139 7.213
2 150-160lb 536 2.297 0.832 0.140 7.136
3 160-170lb 644 1.347 0.298 0.076 13.177
4 170-180lb 567 2.281 0.825 0.067 14.961
5 >180lb 817 2.700 0.993 0.065 15.465
6 >
7 > ##stratum-specific CIs
8 > conf.level <- 0.95
9 > Z <- qnorm(0.5*(1 + conf.level))
10 > Z
11 [1] 1.959964
12 > logrri.L <- log.rri - Z*sqrt(var.log.rri)
13 > logrri.U <- log.rri + Z*sqrt(var.log.rri)
14 > rri.L <- exp(logrri.L)
15 > rri.U <- exp(logrri.U)
16 >
17 > ##create table 2.9, p. 32
18 > tab2.8 <- cbind(ni, log.rri, logrri.L, logrri.U, rri, rri.L, rri.U)
19 > round(tab2.8, 3)
20      ni log.rri logrri.L logrri.U rri rri.L rri.U
21 <150lb 590 0.924 0.194 1.654 2.520 1.215 5.228
22 150-160lb 536 0.832 0.098 1.565 2.297 1.103 4.784
23 160-170lb 644 0.298 -0.242 0.838 1.347 0.785 2.311
24 170-180lb 567 0.825 0.318 1.331 2.281 1.374 3.786
25 >180lb 817 0.993 0.495 1.492 2.700 1.640 4.445
26 >
27 > ##risk ratios seemed approximately equal, but different than 1
28 > ##calculate weighted average rr, p. 33
29 > log.rr.bar <- sum(wi * log.rri)/sum(wi)
30 > log.rr.bar
31 [1] 0.7631178
32 >
33 > zi <- (log.rri - log.rr.bar)/sd.log.rri
34 > zi
35 <150lb 150-160lb 160-170lb 170-180lb >180lb
36 0.4327692 0.1827991 -1.6886492 0.2375570 0.9053572
37 >
38 > ##[Method 1] woolf's chi square test for homogeneity
39 > chi2.w <- sum(zi^2) ##with k-1 d.f.
40 > chi2.w
41 [1] 3.948346
42 >
43 > sum(wi*(log.rri - log.rr.bar)^2)
44 [1] 3.948346
45 >
46 > 1 - pchisq(q = chi2.w, df = length(zi) - 1)
47 [1] 0.4130416
48 > ##equivalent and preferred
49 > pchisq(q = chi2.w, df = length(zi) - 1, lower.tail = FALSE)
50 [1] 0.4130416
51 >
52 > ##woolf's estimate
53 > rr.w <- exp(log.rr.bar)
54 > rr.w

```

```

1 [1] 2.144953
2 >
3 > ##as an exercise do the MH common risk ratio
4 >
5 > ##now test whether rr.w differs from 1 (or log.rr.bar differs from 0)
6 > var.log.rr.bar <- 1/sum(wi)
7 > z <- (log.rr.bar - log(1))/sqrt(var.log.rr.bar)
8 > z
9 [1] 5.809297
10 > ##P(Z>=z)
11 > pnorm(z, lower.tail=FALSE)
12 [1] 3.136778e-09
13 >
14 > ##CI for common risk ratio (woolf estimate), p. 33
15 > conf.level <- 0.95
16 > Z <- qnorm(0.5*(1 + conf.level))
17 > Z
18 [1] 1.959964
19 > var.log.rr.bar <- 1/sum(wi)
20 > logrrbar.L <- log.rr.bar - Z*sqrt(var.log.rr.bar)
21 > logrrbar.L
22 [1] 0.5056541
23 > logrrbar.U <- log.rr.bar + Z*sqrt(var.log.rr.bar)
24 > logrrbar.U
25 [1] 1.020582
26 > rr.w.L <- exp(logrrbar.L)
27 > rr.w.L
28 [1] 1.658070
29 > rr.w.U <- exp(logrrbar.U)
30 > rr.w.U
31 [1] 2.774808
32 >
33 >
34 >
35 >
36 > ##[method 2] Poisson model
37 > ##No interaction model
38 > m1 <- glm(chd ~ beh + factor(wt), weight = count, data = chddat, family = poisson)
39 > summary(m1)
40
41 Call:
42 glm(formula = chd ~ beh + factor(wt), family = poisson, data = chddat,
43     weights = count)
44
45 Deviance Residuals:
46     Min       1Q   Median       3Q      Max
47  -9.814  -6.165   1.175   8.492  11.839
48
49 Coefficients:
50             Estimate Std. Error z value Pr(>|z|)
51 (Intercept) -3.34898    0.19771  -16.939 < 2e-16 ***
52 beh          0.77373    0.13533   5.717 1.08e-08 ***
53 factor(wt)2  0.05557    0.25201   0.221 0.825468
54 factor(wt)3  0.32883    0.22644   1.452 0.146444

```

```

1 factor(wt)4  0.72353    0.21551    3.357 0.000787 ***
2 factor(wt)5  0.51475    0.21008    2.450 0.014277 *
3 ---
4 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
5
6 (Dispersion parameter for poisson family taken to be 1)
7
8     Null deviance: 1288.8  on 19  degrees of freedom
9 Residual deviance: 1233.8  on 14  degrees of freedom
10 AIC: 1759.8
11
12 Number of Fisher Scoring iterations: 6
13
14 >
15 > ##Interaction model
16 > m2 <- glm(chd ~ beh + factor(wt) + beh*factor(wt), weight = count, data = chddat,
17 family = poisson)
18 > summary(m2)
19
20 Call:
21 glm(formula = chd ~ beh + factor(wt) + beh * factor(wt), family = poisson,
22     data = chddat, weights = count)
23
24 Deviance Residuals:
25     Min       1Q   Median       3Q      Max
26 -10.103   -6.223    1.245    8.468   11.585
27
28 Coefficients:
29             Estimate Std. Error z value Pr(>|z|)
30 (Intercept)   -3.44999    0.31623  -10.910 <2e-16 ***
31 beh             0.92426    0.38138   2.423  0.0154 *
32 factor(wt)2    0.11778    0.44721   0.263  0.7923
33 factor(wt)3    0.73246    0.38421   1.906  0.0566 .
34 factor(wt)4    0.78862    0.39068   2.019  0.0435 *
35 factor(wt)5    0.45426    0.39068   1.163  0.2449
36 beh:factor(wt)2 -0.09271    0.54136  -0.171  0.8640
37 beh:factor(wt)3 -0.62633    0.47703  -1.313  0.1892
38 beh:factor(wt)4 -0.09972    0.46836  -0.213  0.8314
39 beh:factor(wt)5  0.06908    0.46372   0.149  0.8816
40 ---
41 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
42
43 (Dispersion parameter for poisson family taken to be 1)
44
45     Null deviance: 1288.8  on 19  degrees of freedom
46 Residual deviance: 1230.2  on 10  degrees of freedom
47 AIC: 1764.2
48
49 Number of Fisher Scoring iterations: 6
50
51 >
52 > ##Test for homogeneity (interaction)
53 > m12.resid.dev <- m1$deviance - m2$deviance
54 > m12.resid.dev

```

```

1 [1] 3.600578
2 > m12.df.resid <- m1$df.residual - m2$df.residual
3 > m12.df.resid
4 [1] 4
5 > pchisq(q = m12.resid.dev, df = m12.df.resid, lower.tail = FALSE)
6 [1] 0.4627509
7 > ##short cut to compare nested models
8 > anova(m1, m2, test = "Chisq")
9 Analysis of Deviance Table
10
11 Model 1: chd ~ beh + factor(wt)
12 Model 2: chd ~ beh + factor(wt) + beh * factor(wt)
13   Resid. Df Resid. Dev Df Deviance P(>|Chi|)
14 1         14    1233.79
15 2          9    1230.19  4      3.60    0.46
16 >
17 >
18 > ##[method 3: pearson chi2 goodness of fit]
19 > ##get marginal totals
20 > mtot <- tapply(chddat$count, list(chddat$beh, chddat$wt), sum)
21 > mtot
22   1  2  3  4  5
23 0 315 280 318 272 380
24 1 275 256 326 295 437
25 > mtot <- as.vector(mtot)
26 > mtot
27 [1] 315 275 280 256 318 326 272 295 380 437
28 > mtot <- rep(mtot, rep(2, length(mtot)))
29 > mtot
30 [1] 315 315 275 275 280 280 256 256 318 318 326 326 272 272 295 295 380 380 437
31 [20] 437
32 > cellprob <- abs(rep(c(1,0), nrow(chddat)/2) - fitted(m1))
33 > cellprob
34   1         2         3         4         5         6         7
35 0.96487978 0.03512022 0.92386498 0.07613502 0.96287278 0.03712722 0.91951414
36   8         9         10        11        12        13        14
37 0.08048586 0.95120598 0.04879402 0.89422239 0.10577761 0.92759289 0.07240711
38  15        16        17        18        19        20
39 0.84303300 0.15696700 0.94123612 0.05876388 0.87260933 0.12739067
40 > ei <- cellprob*mtot
41 >
42 > tab2.11 <- cbind(chddat[,1:3], Total = mtot, pij = cellprob,
43 +                 oi = chddat[,4], ei = ei)
44 > tab2.11
45   chd beh wt Total      pij oi      ei
46 1    0  0  1   315 0.96487978 305 303.93713
47 2    1  0  1   315 0.03512022  10  11.06287
48 3    0  1  1   275 0.92386498 253 254.06287
49 4    1  1  1   275 0.07613502  22  20.93713
50 5    0  0  2   280 0.96287278 270 269.60438
51 6    1  0  2   280 0.03712722  10  10.39562
52 7    0  1  2   256 0.91951414 235 235.39562
53 8    1  1  2   256 0.08048586  21  20.60438
54 9    0  0  3   318 0.95120598 297 302.48350

```

```

1 10 1 0 3 318 0.04879402 21 15.51650
2 11 0 1 3 326 0.89422239 297 291.51650
3 12 1 1 3 326 0.10577761 29 34.48350
4 13 0 0 4 272 0.92759289 253 252.30527
5 14 1 0 4 272 0.07240711 19 19.69473
6 15 0 1 4 295 0.84303300 248 248.69473
7 16 1 1 4 295 0.15696700 47 46.30527
8 17 0 0 5 380 0.94123612 361 357.66972
9 18 1 0 5 380 0.05876388 19 22.33028
10 19 0 1 5 437 0.87260933 378 381.33028
11 20 1 1 5 437 0.12739067 59 55.66972
12 >
13 > ##Pearson chi-square goodness-of-fit statistic, p. 37
14 > chi2 <- sum((tab2.11[,"oi"] - tab2.11[,"ei"])^2/tab2.11[,"ei"])
15 > chi2
16 [1] 3.995289
17 > ##recall that d.f. = (r-1)(c-1)(d-1) for 3-way array
18 > pchisq(chi2, df = 4, lower.tail = FALSE)
19 [1] 0.4066437
20 >

```