Chapter Nine

Ryan Honea

10/9/2017

Exercise Two

Question

The data in the table below (Table 9.13) are numbers of insurance policies, n, and numbers of claims, y, for cars in various insurance categories, CAR, tabublated by age of policy holder, AGE, and district where the policy holder lived ($DIST = 1$, for London and other major cities, and $DIST = 0$, otherwise). The table is derived from the $CLAIMS$ data set in Aitkin et al. (2005) obtained from a paper by Baxter et al. (1980).

		DIST = 0		DIST = 1	
CAR	AGE	y	n	y	n
1	1	65	317	2	20
1	2	65	476	5	33
1	3	52	486	4	40
1	4	310	3259	36	316
2	1	98	486	7	31
2	2	159	1004	10	81
2	3	175	1355	22	122
2	4	877	7660	102	724
3	1	41	223	5	18
3	2	117	539	7	39
3	3	137	697	16	68
3	4	477	3442	63	344
4	1	11	40	0	3
4	2	35	148	6	16
4	3	39	214	8	25
4	4	167	1019	33	114

Solution

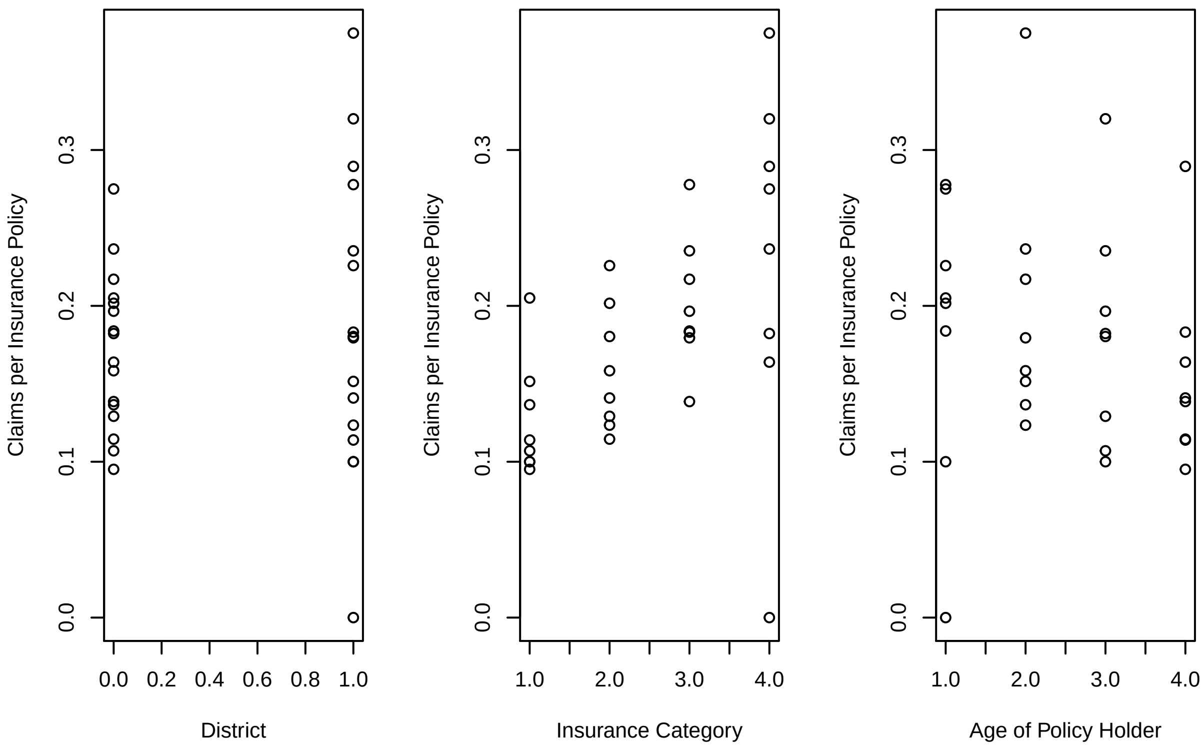
Below is just data entry.

```
df <- data.frame(
  CAR = CAR <- rep(rep(1:4, each = 4),2),
  AGE = AGE <- rep(rep(1:4, 4),2),
  DIST = DIST <- rep(0:1, each = 16),
  y = y <- c(  65,  65,  52, 310,  98, 159, 175, 877,
              41, 117, 137, 477,  11,  35,  39, 167,
               2,   5,   4,  36,   7,  10,  22, 102,
               5,   7,  16, 63,   0,   6,   8,  33),
  n = n <- c( 317, 476, 486,3259, 486,1004,1355,7660,
             223, 539, 697,3442,  40, 148, 214,1019,
              20,  33,  40, 316,  31,  81, 122, 724,
              18,  39,  68, 344,   3,  16,  25, 114)
)
```

(a): Calculate the rate of claims y/n for each category and plot the rates by AGE, CAR, and $DIST$ to get an idea of the main effects of those factors.

Solution:

```
df$rate <- df$y/df$n
par(mfrow = c(1,3))
plot(df$DIST, df$rate, xlab= "District", ylab = "Claims per Insurance Policy")
plot(df$CAR, df$rate, xlab= "Insurance Category", ylab = "Claims per Insurance Policy")
plot(df$AGE, df$rate, xlab= "Age of Policy Holder", ylab = "Claims per Insurance Policy")
```



Based on these plots, it appears that the difference between districts is just an increased variance of rates in lower population areas. Insurance categories from 1 to 4 seem to have increasing rates as well as increasing variance per rate level. From the youngest age group to the second youngest, there is an increase, but a decrease following that in claim rates.

(b): Use Poisson regression to estimate the main effects (each treated as categorical and modeled using indicator variables) and interaction terms.

Solution: First, we convert `CAR`, `AGE`, and `DIST` into factors so that they are treated categorically.

```
df$CAR <- factor(df$CAR); df$AGE <- factor(df$AGE); df$DIST <- factor(df$DIST)
poismod <- glm(y ~ CAR*AGE*DIST + offset(log(n)), data = df,
               family = poisson(link = "log"))
summary(poismod)
```

```
##
## Call:
## glm(formula = y ~ CAR * AGE * DIST + offset(log(n)), family = poisson(link = "log"),
##     data = df)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
## [24]  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.585e+00  1.240e-01 -12.775  < 2e-16 ***
## CAR2            -1.673e-02  1.600e-01  -0.105 0.916721
## CAR3            -1.091e-01  1.994e-01  -0.547 0.584401
## CAR4             2.935e-01  3.260e-01   0.900 0.367947
## AGE2            -4.065e-01  1.754e-01  -2.317 0.020477 *
## AGE3            -6.504e-01  1.860e-01  -3.496 0.000472 ***
## AGE4            -7.681e-01  1.364e-01  -5.630  1.8e-08 ***
## DIST1           -7.181e-01  7.179e-01  -1.000 0.317198
## CAR2:AGE2        1.649e-01  2.174e-01   0.759 0.448106
## CAR3:AGE2        5.726e-01  2.524e-01   2.269 0.023298 *
## CAR4:AGE2        2.556e-01  3.876e-01   0.660 0.509574
## CAR2:AGE3        2.049e-01  2.248e-01   0.912 0.361989
## CAR3:AGE3        7.173e-01  2.575e-01   2.785 0.005345 **
## CAR4:AGE3        2.390e-01  3.888e-01   0.615 0.538711
## CAR2:AGE4        2.021e-01  1.731e-01   1.168 0.242996
## CAR3:AGE4        4.854e-01  2.124e-01   2.286 0.022271 *
## CAR4:AGE4        2.505e-01  3.399e-01   0.737 0.461104
## CAR2:DIST1       8.312e-01  8.176e-01   1.017 0.309299
## CAR3:DIST1       1.131e+00  8.601e-01   1.315 0.188626
## CAR4:DIST1      -2.139e+01  4.225e+04  -0.001 0.999596
## AGE2:DIST1       8.220e-01  8.548e-01   0.962 0.336246
## AGE3:DIST1       6.504e-01  8.858e-01   0.734 0.462753
## AGE4:DIST1       8.984e-01  7.392e-01   1.215 0.224188
## CAR2:AGE2:DIST1 -1.184e+00  9.950e-01  -1.190 0.234003
## CAR3:AGE2:DIST1 -1.425e+00  1.052e+00  -1.354 0.175588
## CAR4:AGE2:DIST1  2.175e+01  4.225e+04   0.001 0.999589
## CAR2:AGE3:DIST1 -4.298e-01  9.944e-01  -0.432 0.665567
## CAR3:AGE3:DIST1 -8.832e-01  1.039e+00  -0.850 0.395125
## CAR4:AGE3:DIST1  2.202e+01  4.225e+04   0.001 0.999584
## CAR2:AGE4:DIST1 -8.042e-01  8.428e-01  -0.954 0.340026
## CAR3:AGE4:DIST1 -1.032e+00  8.881e-01  -1.162 0.245079
## CAR4:AGE4:DIST1  2.178e+01  4.225e+04   0.001 0.999589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2.0783e+02  on 31  degrees of freedom
## Residual deviance: 4.1219e-10  on  0  degrees of freedom
## AIC: 232.36
##
## Number of Fisher Scoring iterations: 20
```

(c): Based on the modelling in (b), Aitkin et al. (2005) determined that all the interactions were unimportant and decided that AGE and CAR could could be treated as though they were continuous variables. Fit a model incorpororating these features and compare it with the best model in (b). What conclusions do you reach?

Solution: First, we convert `CAR` and `AGE` into numeric values and then create the other model.

```r
df$CAR <- as.numeric(df$CAR); df$AGE <- as.numeric(df$AGE)
poismod2 <- glm(y ~ CAR + AGE + DIST + offset(log(n)), data = df,
                family = poisson(link = "log"))
summary(poismod2)
```

```
##
## Call:
## glm(formula = y ~ CAR + AGE + DIST + offset(log(n)), family = poisson(link = "log"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7248  -0.5681  -0.1679   0.3384   1.9126
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.85253    0.07990 -23.185  < 2e-16 ***
## CAR          0.19777    0.02080   9.507  < 2e-16 ***
## AGE         -0.17674    0.01849  -9.559  < 2e-16 ***
## DIST1        0.21865    0.05853   3.736 0.000187 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 207.833  on 31  degrees of freedom
## Residual deviance:  24.685  on 28  degrees of freedom
## AIC: 201.05
##
## Number of Fisher Scoring iterations: 4
```

We can check whether or not this model fits well by comparing it to the saturated model which was found in (b). This is the residual deviance noted in the summary.

```r
p <- pchisq(poismod2$deviance, 28)
cat("p-value: ", p)
```

```
## p-value:  0.3550971
```

Based on the p-value, there is not enough evidence to suggest that this model does not fit the data and so we would prefer to use this simpler model.

Exercise Three

Question

This question relates to the flu vaccine trial data in the table below (Table 9.6).

	Response			
	Small	Moderate	Large	Total
Placebo	25	8	5	38
Vaccine	6	18	11	35

Solution

Below is data entry.

```
vacctrial <- as.table( matrix(
  c(25, 8, 5, 6, 18, 11), byrow = TRUE, ncol = 3)
)
colnames(vacctrial) <- c("Small", "Moderate", "Large")
rownames(vacctrial) <- c("Vaccine", "Placebo")
```

(a): Using a conventional chi-squared test and an appropriate log-linear model, test the hypothesis that the distribution of responses is the same for the placebo and vaccine groups.

Solution: The chi-squared test is performed below.

```
require(MASS)
```

```
## Loading required package: MASS
```

```
chisq.test(vacctrial)
```

```
##
##  Pearson's Chi-squared test
##
## data:  vacctrial
## X-squared = 17.648, df = 2, p-value = 0.0001472
```

With at least 95% confidence, we reject the null hypothessis that the distribution of the responses are the same for the placebo and vaccine groups. We test this again using the additive log-linear model.

```
df <- as.data.frame(vacctrial)
colnames(df) <- c("Treatment", "Response", "Freq")
df$Treatment <- relevel(df$Treatment, ref = "Placebo")
df$Response <- relevel(df$Response, ref = "Small")
poismod <- glm(Freq ~ Treatment + Response, data = df,
               family = poisson(link = "log"))
fits <- c(fitted.values(poismod))
X2 <- sum(((df$Freq - fits)/sqrt(fits))^2)
1 - pchisq(X2, 2)
```

```
## [1] 0.0001471709
```

With an equal p-value, we make the same conclusion.

(b): For the model corresponding to the hypothesis of homogeneity of response distributions, calculate the fitted values, the Pearson and deviance residuals, and the goodness of fit statistics χ^2 and D. Which of the cells of the table contribute most to χ^2 (or D)? Explain and interpret these results.

Solution: We test the hypothesis of homogeneity by comparing the saturated and the additive model.

```
poissat <- glm(Freq ~ Treatment*Response, data = df,
               family = poisson(link = "log"))
dev.res <- sign(df$Freq - fits) *(sqrt(2*(df$Freq*log(df$Freq/fits)-(df$Freq-fits))))
deviance <- sum(dev.res^2)
deviance
```

```
## [1] 18.64253
```

```
2*(logLik(poissat) - logLik(poismod))
```

```
## 'log Lik.' 18.64253 (df=6)
```

These give the same results as before and so we see homogeneity of response distribution.

(c): Re-analyze these data using ordinal logistic regression to estimate cut-points for for a latent continuous response variable and to estimate a location shift between the two treatment groups. Sketch a rough diagram to illustrate the model which forms the conceptual base for this analysis (see Exercise 8.4).

Solution:

```
require(nnet)
```

```
## Loading required package: nnet
```

```
ordmod <- polr(Response ~ Treatment, weights = Freq, data = df)
df$probs <- predict(ordmod, type = "probs")
df[1:2,c("Treatment", "probs")]
```

```
##   Treatment probs.Small probs.Moderate probs.Large
## 1   Vaccine  0.63761389     0.28226555  0.08012056
## 2   Placebo  0.21883671     0.42755801  0.35360528
```

```
shift <- log((df[2,4][1]*(df[1,4][2]+df[1,4][3]))/(df[1,4][1]*(df[2,4][2]+df[2,4][3])))
cat("Estimated Shift:", shift)
```

```
## Estimated Shift: -1.837481
```