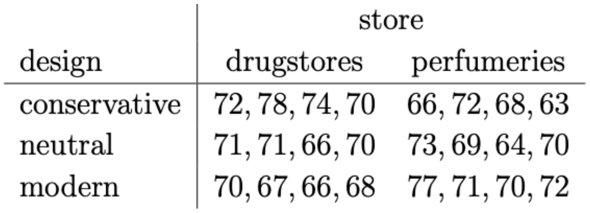
# Exercises05

Dennis Perrett

1/14/2022

***Exercise 18.*** The marketing department of a cosmetics company is conducting an expe- riment to study the effect of alternative design variants of a product on the sales. Sales (in units per 1000 customers) are monitored in 24 randomly selected stores (either (1) drugstores or (2) perfumeries). The design variants are (1) conservative, (2) neutral and (3) modern. The following data on the sales are available:



| | store | |
|---|---|---|
| design | drugstores | perfumeries |
| conservative | 72, 78, 74, 70 | 66, 72, 68, 63 |
| neutral | 71, 71, 66, 70 | 73, 69, 64, 70 |
| modern | 70, 67, 66, 68 | 77, 71, 70, 72 |

.

```
##
## Call:
## lm(formula = sales ~ design * store, data = dt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.000 -1.562  0.125  1.500  4.750
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     73.500      1.557  47.212  < 2e-16 ***
## designmodern                    -5.750      2.202  -2.612  0.01766 *
## designneutral                   -4.000      2.202  -1.817  0.08593 .
## storeperfumaries                -6.250      2.202  -2.839  0.01089 *
## designmodern:storeperfumaries   11.000      3.114   3.533  0.00238 **
## designneutral:storeperfumaries   5.750      3.114   1.847  0.08129 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.114 on 18 degrees of freedom
## Multiple R-squared:  0.4257, Adjusted R-squared:  0.2661
## F-statistic: 2.668 on 5 and 18 DF,  p-value: 0.05655
```

```
## Analysis of Variance Table
##
## Response: sales
##              Df  Sum Sq Mean Sq F value  Pr(>F)
## design        2   5.583   2.792  0.2880 0.75318
## store         1   2.667   2.667  0.2751 0.60635
## design:store  2 121.083  60.542  6.2450 0.00871 **
## Residuals    18 174.500   9.694
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpret these results by answering the following questions:

a. Which statistical model was used in analysing the data?

- linear regression?

b. Specify the hypothesis that may be investigated.

- $\beta_0 = \beta_1 = \beta_i = 0$

c. Which value is to be checked, or which values are to be checked in order to decide whether there is an effect of product design on sales?

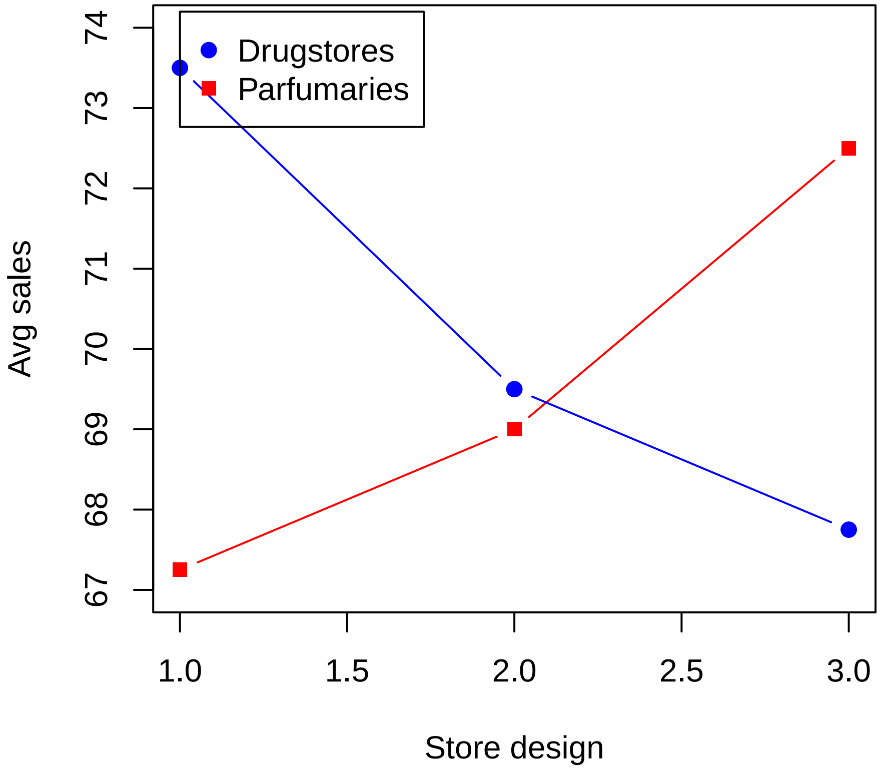
- design:store in the anova table. We look at this value, because the interaction term renders the base term (design) uninterpretable.

d. What do you conclude on this effect ($\alpha = 0.05$)?

- P-values etc associated with design:store is significant (0.00871) at the 5% level. From this we can conclude that the type of store design has an effect on sales.

e. Sketch the results of the experiment (abscissa: design variants; ordinate: average sales per group) and explain how an interaction effect is (or, would be) represented graphically.

- The cross over between the 2 lines indicates significance. An interaction effect would be represented by a new axis.



f. Which marketing strategy is (i.e. which product design in which store) would you recommend based on the results?

- It depends on the store type. This is visible in the graph. Store design 3 (modern) performs best for parfumaries. Stores design 1 (conservative) performs best for drug stores.

g. The analysis is based on the general linear model

$$Y_{ijk} = \mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

with independent and normally distributed errors $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$, and $\mu_0$ the mean sales (total expected value). Provide estimates for the parameters $\mu_0, \alpha_i, \beta_j, (\alpha\beta)_{ij}$, and $\sigma^2$.

$\mu_0 = 73.5$ Overall intercept. (In original case 69.9)

$\alpha_i = (-5.750, -4.000, 9.750)$ The change dependent on design type.
(Original example (0.458, -0.6667, 0.2087))

$\beta = (-6.250, 6.250)$ Shift in intercept based on store type.
(Original case (0.333,-0.333))

$$(\alpha\beta)_{ij} = \begin{pmatrix} 11.000 & -11.000 \\ 5.750 & -5.750 \\ 16.750 & -16.750 \end{pmatrix}$$

Interaction term ceofficients. Rows and columns must sum to 0.

Original exercise: $\begin{pmatrix} 2.79 & -2.79 \\ -0.083 & 0.083 \\ -2.708 & 2.708 \end{pmatrix}$

$\sigma^2 = 3.114^2 = 9.6969$ Squared Residual deviance.

**Exercise 19.** In the context of a methadone program, a study by Anglin, McGlothin & Speckart (1981) examined a sample of 647 heroin addicts. The following variables were collected:

| | |
|---|---|
| $C$ | County of residence: urban, suburban, rural |
| $S$ | Socioeconomic status (low, not low) |
| $P$ | Parents available till 16: yes, no |

The dependent variable F indicates whether or not the first heroin use was before the 18th birthday. The following frequencies were observed:

```
##
## Call:
## glm(formula = cbind(yes, no) ~ S + C + P + C * P, family = binomial,
##     data = dat)
##
## Deviance Residuals:
##        1         3         5         7         9        11        13        15
## -0.01538   0.61443  -0.74845   0.01421  -0.63023   0.98281  -0.63977   0.67228
##       17        19        21        23
##  0.59013   0.63079  -1.17908  -0.95387
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.7053     0.1733  -4.070 4.71e-05 ***
## Snot           -0.3459     0.1717  -2.014 0.044019 *
## Csuburban       0.5621     0.2670   2.105 0.035250 *
## Crural         -0.4229     0.3073  -1.376 0.168709
## Pno             0.8217     0.2130   3.858 0.000114 ***
## Csuburban:Pno  -1.0938     0.4157  -2.631 0.008505 **
## Crural:Pno     -0.7637     0.5516  -1.384 0.166255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 37.1726  on 11  degrees of freedom
## Residual deviance:  6.2088  on  5  degrees of freedom
## AIC: 67.929
##
## Number of Fisher Scoring iterations: 4
```

Calculate by hand and with R

a. Specify the underlying design matrix X. Arrange columns according to the sequence of parameter estimates in the above table.

$$X = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

```
model.matrix(lm2)
```

```
##    (Intercept) Snot Csuburban Crural Pno Csuburban:Pno Crural:Pno
## 1            1    0         0      0   0             0          0
## 3            1    0         1      0   0             0          0
## 5            1    0         0      1   0             0          0
## 7            1    1         0      0   0             0          0
## 9            1    1         1      0   0             0          0
## 11           1    1         0      1   0             0          0
## 13           1    0         0      0   1             0          0
## 15           1    0         1      0   1             1          0
## 17           1    0         0      1   1             0          1
## 19           1    1         0      0   1             0          0
## 21           1    1         1      0   1             1          0
## 23           1    1         0      1   1             0          1
## attr(,"assign")
## [1] 0 1 2 2 3 4 4
## attr(,"contrasts")
## attr(,"contrasts")$S
## [1] "contr.treatment"
##
## attr(,"contrasts")$C
## [1] "contr.treatment"
##
## attr(,"contrasts")$P
## [1] "contr.treatment"
```

b. Determine the data vector $\vec{x}$ of a heroin addict who grew up without parents in the suburbs, and has low socioeconomic status.

- $\vec{x} = (1, 1, 0, 1, 0, 0, 1)$

c. What are the odds and the probability predicted by the model for this person?

- Log-Odds: $(-1.416 + 0.34585 + 0 + 0.65490 + 0 + 0 + 0.33017 = -0.4153013$
- Odds: $e^{Log-Odds} = e^{-0.4153013} = 0.6601413$
- Probability: $\frac{1}{1+e^{-Log-Odds}} = \frac{1}{1+e^{0.4153013}} = 0.3976416$

```
x = matrix(c(1,1,0,1,0,0,0),1,7)
log.odds <- x%*%lm2$coefficients
odds <- exp(log.odds)
prob <- 1/(1+exp(-log.odds))
```

d. Compute the odds for a person who, all other things being equal (ceteris paribus), has no low socioeconomic status. Compute the odds ratio of the previously considered person compared to this person?

- Log-Odds: $(-1.416 + 0 + 0 + 0.65490 + 0 + 0 + 0.33017) = -0.7611535$
- Odds: $e^{Log-Odds} = e^{-0.7611535} = 0.4671273$
- Probability: $\frac{1}{1+e^{-Log-Odds}} = \frac{1}{1+e^{0.7611535}} = 0.3183959$
- Odds ratio: $\frac{e^{-1.41606} \cdot e^{0.34585} \cdot e^{0.33017} \cdot e^{0.65490}}{e^{-1.41606} \cdot e^{0.33017} \cdot e^{0.65490}} = e^{0.34585} = 1.413191$

```
x = matrix(c(1,0,0,1,0,0,0),1,7)
log.odds <- x%*%lm2$coefficients
odds <- exp(log.odds)
prob <- 1/(1+exp(-log.odds))
```

e. How does the result of the preceding computation relate to the estimated value of the parameter associated with S *low* in the table?

- Normally we cannot interpret non-interaction terms, when interaction terms are included. In this case, the interaction term(s) are all 0. As such, the rest of the terms are interpretable. The odds are 1.41 times greater for someone with low socio-economic status (ceteris paribus) compared to someone with high socio-economic status.

f. Interpret the effect of the availability of the parents by age 16 on the time of first heroin use according to the model.

- $exp(-0.5805) = 0.944$. This is close to one. Availability has little effect as odds are only 0.944 times.

g. Decide whether the data are appropriately described by the model ($\alpha = 0.05$). (Hint: $\chi^2_{0.95}(5) = 11.07$).

- $H_0$: Saturated model does not describe the data better than the given model.
- Residual variance = 6.2088. $6.2088 < 11.07$. Cannot reject null: saturated model is not better than given model. Model is good.

**Exercise 20.** Costa et al. (2014) report data of experiments that they interpret as evidence for a so-called 'foreign language effect on framing'. This involves the preference for a positively framed (presented as a gain) over a negatively framed (presented as a loss) safe option over a risky option in a decision task, which is supposed to be less pronounced when presented in a foreign language compared to the respective native language. In one of the experiments they obtain the following data (Costa et al., 2014, Table 1, Experiment AD1):

```
##   safe risky       L   Fr
## 1   42    20  native gain
## 2   21    41  native loss
## 3   41    20 foreign gain
## 4   31    31 foreign loss
```

Use R (Type I error $\alpha = 0.05$) to show:

a. that a significant result in terms of a framing effect results for the 2 × 2 table for the native language, whereas this is not the case for the 2 × 2 table for the foreign language,

```
glm2 <- glm(cbind(safe, risky) ~ Fr,binomial, dat2[dat2$L == "native",])
glm3 <- glm(cbind(safe, risky) ~ Fr, binomial, dat2[dat2$L == "foreign",])
glm4 <- glm(cbind(safe, risky) ~ Fr + L + Fr*L, binomial, dat2)

summary(glm2) # Gain is significant
```

```
##
## Call:
## glm(formula = cbind(safe, risky) ~ Fr, family = binomial, data = dat2[dat2$L ==
##     "native", ])
##
## Deviance Residuals:
## [1]  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.7419     0.2717   2.731  0.00632 **
## Frloss       -1.4110     0.3819  -3.695  0.00022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.4515e+01  on 1  degrees of freedom
## Residual deviance: 9.3259e-15  on 0  degrees of freedom
## AIC: 12.932
##
## Number of Fisher Scoring iterations: 3
```

```
summary(glm3) # Gain is insignificant
```

```
##
## Call:
## glm(formula = cbind(safe, risky) ~ Fr, family = binomial, data = dat2[dat2$L ==
##     "foreign", ])
##
## Deviance Residuals:
## [1]  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.7178     0.2727   2.632  0.00849 **
## Frloss       -0.7178     0.3727  -1.926  0.05410 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3.7766e+00  on 1  degrees of freedom
## Residual deviance: 8.6597e-15  on 0  degrees of freedom
## AIC: 13.033
##
```

```
## Number of Fisher Scoring iterations: 3
```

b. that, however, a 'foreign language effect on framing' cannot be demonstrated.

```
summary(glm4) # Framing effect is insignificant
```

```
##
## Call:
## glm(formula = cbind(safe, risky) ~ Fr + L + Fr * L, family = binomial,
##     data = dat2)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.7178     0.2727   2.632  0.00849 **
## Frloss          -0.7178     0.3727  -1.926  0.05410 .
## Lnative          0.0241     0.3850   0.063  0.95009
## Frloss:Lnative  -0.6932     0.5336  -1.299  0.19394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.9782e+01  on 3  degrees of freedom
## Residual deviance: 3.9524e-14  on 0  degrees of freedom
## AIC: 25.965
##
## Number of Fisher Scoring iterations: 3
```

(Note: In two other experiments, basically the same picture emerges, so that an integrative evaluation of the data across all three experiments arrives at the same conclusion).

Alternative solutions:

```
chisq.test(as.matrix(dat2[dat2$L == "native",1:2]),correct=F) # Significant
```

```
##
##  Pearson's Chi-squared test
##
## data:  as.matrix(dat2[dat2$L == "native", 1:2])
## X-squared = 14.23, df = 1, p-value = 0.0001618
```

```
chisq.test(as.matrix(dat2[dat2$L == "foreign",1:2]),correct=F) # Insignificant
```

```
##
##  Pearson's Chi-squared test
##
## data:  as.matrix(dat2[dat2$L == "foreign", 1:2])
## X-squared = 3.7536, df = 1, p-value = 0.0527
```

```
glm5 <- glm(cbind(safe, risky) ~ Fr + L, binomial, dat2)
glm6 <- glm(cbind(safe, risky) ~ Fr * L, binomial, dat2)

summary(glm6)
```

```
##
```

```
## Call:
## glm(formula = cbind(safe, risky) ~ Fr * L, family = binomial,
##     data = dat2)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.7178     0.2727   2.632  0.00849 **
## Frloss           -0.7178     0.3727  -1.926  0.05410 .
## Lnative           0.0241     0.3850   0.063  0.95009
## Frloss:Lnative   -0.6932     0.5336  -1.299  0.19394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.9782e+01  on 3  degrees of freedom
## Residual deviance: 3.9524e-14  on 0  degrees of freedom
## AIC: 25.965
##
## Number of Fisher Scoring iterations: 3
```

```
anova(glm5, glm6, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(safe, risky) ~ Fr + L
## Model 2: cbind(safe, risky) ~ Fr * L
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         1     1.6929
## 2         0     0.0000  1   1.6929   0.1932
```