

## Solution to Series 7

```
1. a) > count <- c(31,28,33,38,28,32,39,27,28,39,21,39,45,37,
                  41,14,16,18,9,21,21,14,12,13,13,14,20,24,
                  15,24,18,13,19,14,15,16,14,19,25,16,16,18,9,10,9)
> probe <- factor(rep(1:3, each = 15))
> vol <- c(rep(40,15),rep(20,30))
> nema <- data.frame(probe,count,vol)
> mod1 <- glm(count~probe,family=poisson,data=nema)
> summary(mod1)
```

Call:

```
glm(formula = count ~ probe, family = poisson, data = nema)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.3580	-0.9031	-0.1267	0.8846	2.2417

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	3.51849	0.04446	79.146	<2e-16 ***
probe2	-0.71311	0.07751	-9.200	<2e-16 ***
probe3	-0.78412	0.07941	-9.875	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 188.602 on 44 degrees of freedom
Residual deviance: 52.528 on 42 degrees of freedom
AIC: 276.14
```

Number of Fisher Scoring iterations: 4

```
> anova(mod1)
```

Analysis of Deviance Table

Model: poisson, link: log

Response: count

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev
NULL			44	188.602
probe	2	136.07	42	52.528

b) There is a large difference between probe 1 and the other two. However, probe 1 has a different concentration which could account for the difference discovered.

```
c) > mod2 <- glm(count~log(vol),family=poisson,data=nema)
> summary(mod2)
```

```
Call:
glm(formula = count ~ log(vol), family = poisson, data = nema)
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-2.3580	-0.7674	-0.1267	0.7368	2.0861

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.46223	0.30991	-1.491	0.136
log(vol)	1.07911	0.09197	11.733	<2e-16 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)
```

```
Null deviance: 188.602 on 44 degrees of freedom
```

```
Residual deviance: 53.131 on 43 degrees of freedom
```

```
AIC: 274.74
```

```
Number of Fisher Scoring iterations: 4
```

```
> anova(mod2)
```

```
Analysis of Deviance Table
```

```
Model: poisson, link: log
```

```
Response: count
```

```
Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev
NULL			44	188.602
log(vol)	1	135.47	43	53.131

```
d) > confint(mod2)
```

	2.5 %	97.5 %
(Intercept)	-1.0721154	0.1430996
log(vol)	0.8988966	1.2595331

```
The confidence interval for  $\beta_1$  includes 1.
```

```
The model  $\lambda_i = cvol_i$  is appropriate.
```

```
e) > mod3 <- glm(count~offset(log(vol)),family=poisson,data=nema)
> summary(mod3)
```

```
Call:
```

```
glm(formula = count ~ offset(log(vol)), family = poisson, data = nema)
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-2.2127	-0.8656	-0.1033	0.8548	2.0091

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.19744	0.03186	-6.196	5.78e-10 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 53.871 on 44 degrees of freedom  
Residual deviance: 53.871 on 44 degrees of freedom  
AIC: 273.48

Number of Fisher Scoring iterations: 4

```
> anova(mod3)
```

Analysis of Deviance Table

Model: poisson, link: log

Response: count

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev
NULL			44	53.871

The model with estimated coefficient for log(vol) shows only minor difference to the offset model.

```
2. a) > library(foreign, lib=lib)
> pension <- read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge2k/PENSION.DTA")
> pension$pctstck <- ordered(pension$pctstck)
> pension$choice <- factor(pension$choice)
> pension$female <- factor(pension$female)
> pension$married <- factor(pension$married)
> pension$black <- factor(pension$black)
> pension$prftshr <- factor(pension$prftshr)
```

```
> table(pension$choice,pension$pctstck)
      0 50 100
0 35 28 24
1 43 57 39
> prop.table(table(pension$choice,pension$pctstck),1)
      0      50      100
0 0.4022989 0.3218391 0.2758621
1 0.3093525 0.4100719 0.2805755
```

People with freedom to choose their investment strategy avoid portfolios mainly consisting of obligations.

```
b/c) > pension$inc <- rep(1,226)
> pension$inc[pension$finc35==1 | pension$finc50==1] <- 2
> pension$inc[pension$finc75==1 | pension$finc100==1 | pension$finc101==1] <- 3
> pension$inc <- factor(pension$inc,labels=c("<=25'000", "25'001 to 50'000", "above 50'000"))
> table(pension$inc,pension$pctstck)
      0 50 100
<=25'000      31 15 20
25'001 to 50'000 28 37 28
above 50'000      19 33 15
> prop.table(table(pension$inc,pension$pctstck),1)
```

	0	50	100
<=25'000	0.4696970	0.2272727	0.3030303
25'001 to 50'000	0.3010753	0.3978495	0.3010753
above 50'000	0.2835821	0.4925373	0.2238806

People with a higher income tend to have mixed investment strategies.

```
d) > library(nnet)
> pension$pct <- factor(pension$pctstck, levels = c("50","0","100"),
                        ordered = FALSE)
> mod1 <- multinom(pct~choice+age+educ+female+married+black+inc+wealth89+prftshr,
                  data=pension)

# weights: 36 (22 variable)
initial value 220.821070
iter 10 value 203.476730
iter 20 value 200.261454
iter 30 value 200.186637
final value 200.186632
converged

> summary(mod1)

Call:
multinom(formula = pct ~ choice + age + educ + female + married +
          black + inc + wealth89 + prftshr, data = pension)

Coefficients:
      (Intercept)    choice1         age         educ
0      -2.614677 -0.5317628 0.10229894 -0.1775690
100     1.021584  0.1318421 0.01063465 -0.1168254
      female1    married1      black1 inc25'001 to 50'000
0     -0.172714595 -0.4612883 -0.27305822          -1.0206500
100  -0.006320096 -0.4605590 -0.02921608          -0.3535253
      incabove 50'000    wealth89  prftshr1
0          -0.7282016 0.0006098428 0.1954679
100         -0.6683600 0.0004014558 1.2596317

Std. Errors:
      (Intercept)    choice1         age         educ    female1
0          1.821215 0.3899706 0.03107212 0.07476118 0.4137560
100         1.610395 0.4039064 0.02943977 0.07565837 0.4186522
      married1    black1 inc25'001 to 50'000 incabove 50'000
0    0.5151725 0.6168527          0.4811679          0.5729191
100 0.5066545 0.6001433          0.4859831          0.5968045
      wealth89  prftshr1
0    0.0007823479 0.5087600
100 0.0008517805 0.4759613

Residual Deviance: 400.3733
AIC: 444.3733

e) > mod2 <- multinom(pct~age+educ+female+married+black+inc+wealth89+prftshr,
                    data=pension)

# weights: 33 (20 variable)
initial value 220.821070
iter 10 value 205.380583
iter 20 value 201.836179
final value 201.771474
converged

> deviance(mod2) - deviance(mod1)
```

```
[1] 3.169684
```

```
> anova(mod1, mod2)
```

```

                                Model
1      age + educ + female + married + black + inc + wealth89 + prftshr
2 choice + age + educ + female + married + black + inc + wealth89 + prftshr
  Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
1      382   403.5429      NA      NA      NA
2      380   400.3733 1 vs 2    2  3.169684 0.2049802

```

choice is not significant.

The odds for mainly obligations versus mixed strategy are 1.7 ( $\exp(0.53)$ ) times larger without choice than having a choice.

The odds for mainly stock versus mixed strategy are slightly higher ( $1.14 = \exp(0.13)$ ) when having a choice.

```
f) > predict(mod1, type="probs", newdata=data.frame(choice="0", age=60, educ=13.5, female="0", married=
      50      0      100
0.1934054 0.3954802 0.4111145
> predict(mod1, type="probs", newdata=data.frame(choice="1", age=60, educ=13.5, female="0", married=
      50      0      100
0.2161367 0.2596827 0.5241806
```

```
3. a) > car <- read.table("http://stat.ethz.ch/Teaching/Datasets/car.dat", header=T)
> glm2 <- glm(purchase~income + age, data=car, family=binomial)
> summary(glm2)
```

Call:

```
glm(formula = purchase ~ income + age, family = binomial, data = car)
```

Deviance Residuals:

```

      Min       1Q   Median       3Q      Max
-1.6189  -0.8949  -0.5880   0.9653   2.0846

```

Coefficients:

```

              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.73931     2.10195  -2.255  0.0242 *
income       0.06773     0.02806   2.414  0.0158 *
age          0.59863     0.39007   1.535  0.1249
---

```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

```

Null deviance: 44.987  on 32  degrees of freedom
Residual deviance: 36.690  on 30  degrees of freedom
AIC: 42.69

```

Number of Fisher Scoring iterations: 4

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -4.74 + 0.068 \cdot \text{income} + 0.599 \cdot \text{age}.$$

b)  $\exp \hat{\beta}_{\text{income}} = 1.07$  und  $\exp \hat{\beta}_{\text{age}} = 1.82$ . The odds for buying a new one increase by 7% for each step increase of income by 1000 US \$ and by 82% for each additional year of age of the car.

```
c) > predict(glm2, data.frame(age=3, income=50), type="response")
```

```

1
0.6090245
d) > par(mfrow=c(2,3))
> scatter.smooth(x=fitted(glm2),y=resid(glm2,type="pearson"), span=2/3,degree=1,family="gauss")
> abline(h=0,lty=2)
> scatter.smooth(x=fitted(glm2),y=resid(glm2,type="deviance"), span=2/3,degree=1,family="gauss")
> abline(h=0,lty=2)
> plot(resid(glm2,type="deviance"),ylab="Deviance Residuals")
> hi <- lm.influence(glm2)$hat
> plot(hi,resid(glm2),xlab="leverages",ylab="Deviance Residuals",pch=16,cex=0.8)
> di <- (hi*(resid(glm2,type="pearson")^2))/((glm2$df.null+1-glm2$df.residual)*(1-hi))
> plot(di,pch=16,cex=0.8,las=1, ylab="Cook's Distances")
> identify(di)

```

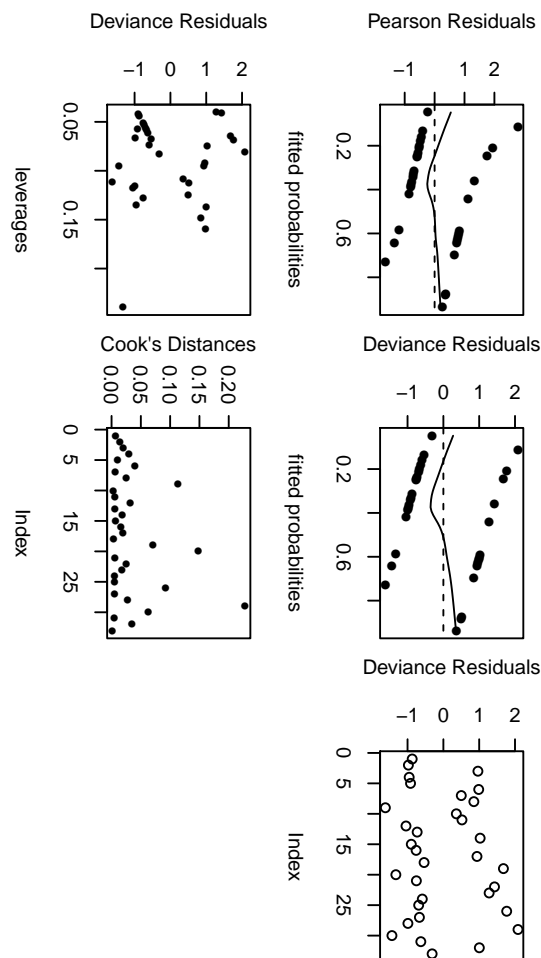


Figure 1: Residual Analysis for Exercise 2.

there seem to be no outliers nor leverage points.

```

e) > glm3 <- glm(purchase~income, data=car, family=binomial)
> (an32 <- anova(glm3,glm2,test="Chisq"))

```

Analysis of Deviance Table

Model 1: purchase ~ income

Model 2: purchase ~ income + age

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	31	39.305			
2	30	36.690	1	2.6149	0.1059

The p-value of 0.106 is larger than 0.05 but still relatively small. It is common practice to be rather lenient with the inclusion of variables in such a situation. The bound to accept a variable can be 0.15 or even 0.20. Thus, we would leave `age` in the model.

```
f) > glm4=glm(purchase~income + age + income:age, data=car, family=binomial)
> summary(glm4)
> anova(glm2,glm4,test="Chisq")
```

Analysis of Deviance Table

```
Model 1: purchase ~ income + age
Model 2: purchase ~ income + age + income:age
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         30      36.690
2         29      35.404  1   1.2855  0.2569
```

there seems to be no interaction between `income` and `age`.

#### 4. Logistic Regression for Binomial Data

In this task we analyze the example concerning hypertension from Altman (1991). First, we need to enter the data. This is done as follows:

```
> no.yes <- c("No", "Yes")
> smoking <- gl(2,1,8, no.yes)
> obesity <- gl(2,2,8, no.yes)
> snoring <- gl(2,4,8, no.yes)
> n.total <- c(60, 17, 8, 2, 187, 85, 51, 23)
> n.hyper <- c(5, 2, 1, 0, 35, 13, 15, 8)
```

Here, the function `gl` creates a factor with given levels. The factors `smoking`, `obesity` and `snoring` have an obvious meaning. `n.total` is the number of observations and `n.hyper` is the number of people with hypertension in each group.

- a) In order to fit a binomial logistic regression model construct a response matrix with two columns containing the number of people with and without hypertension, respectively.

```
> hyper.tbl <- cbind(n.hyper=n.hyper, n.nohyper=n.total-n.hyper)
```

- b) Fit a binomial logistic regression model to the data.

```
> glm.hyp <- glm(hyper.tbl ~ smoking+obesity+snoring, binomial)
```

Here, we model the expected number of people with/without hypertension as a function of the factors `smoking`, `obesity` and `snoring`.

- c) Does this model fit well? Assess the goodness-of-fit via the residual deviance. We perform a chi-squared-test to assess the goodness-of-fit.

```
> pchisq(deviance(glm.hyp), df.residual(glm.hyp), lower=FALSE)
[1] 0.8054809
```

Since this value is way above 0.05 we deduce that this model fits well.

- d) Which variables significantly influence the occurrence of hypertension?

```
> summary(glm.hyp)
```

Call:

```
glm(formula = hyper.tbl ~ smoking + obesity + snoring, family = binomial)
```

Deviance Residuals:

```
      1      2      3      4      5      6
-0.04344  0.54145 -0.25476 -0.80051  0.19759 -0.46602
      7      8
-0.21262  0.56231
```

Coefficients:

```

                Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.37766    0.38018  -6.254  4e-10 ***
smokingYes  -0.06777    0.27812  -0.244  0.8075
obesityYes   0.69531    0.28509   2.439  0.0147 *
snoringYes   0.87194    0.39757   2.193  0.0283 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

(Dispersion parameter for binomial family taken to be 1)

```

Null deviance: 14.1259 on 7 degrees of freedom
Residual deviance: 1.6184 on 4 degrees of freedom
AIC: 34.537

```

Number of Fisher Scoring iterations: 4

From the summary we see that only smoking does not have a significant influence on the response.

e) Try to find a suitable model. Perform likelihood-ratio tests to achieve this goal.

```
> drop1(glm.hyp, test="Chisq")
```

Single term deletions

```

Model:
hyper.tbl ~ smoking + obesity + snoring
      Df Deviance   AIC   LRT Pr(>Chi)
<none>      1.6184 34.537
smoking  1   1.6781 32.597 0.0597  0.80694
obesity  1   7.2750 38.194 5.6566  0.01739 *
snoring  1   7.2963 38.215 5.6779  0.01718 *
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the summary of the regression and the output of drop1 we see that we can exclude smoking from the model.

```
> glm.hyp2 <- glm(hyper.tbl ~ obesity+snoring, binomial)
> summary(glm.hyp2)
```

Call:

```
glm(formula = hyper.tbl ~ obesity + snoring, family = binomial)
```

Deviance Residuals:

```

      1      2      3      4      5      6
-0.01247  0.47756 -0.24050 -0.82050  0.30794 -0.62742
      7      8
-0.14449  0.45770

```

Coefficients:

```

                Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.3921    0.3757  -6.366 1.94e-10 ***
obesityYes   0.6954    0.2851   2.440  0.0147 *
snoringYes   0.8655    0.3967   2.182  0.0291 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

(Dispersion parameter for binomial family taken to be 1)

```

Null deviance: 14.1259 on 7 degrees of freedom
Residual deviance: 1.6781 on 5 degrees of freedom
AIC: 32.597

```

Number of Fisher Scoring iterations: 4



- f) Compare the observed and fitted proportions for hypertension under model e). What is striking here? Additionally, calculate the expected and observed counts.

```
> fitted(glm.hyp2)
      1      2      3      4      5
0.08377892 0.08377892 0.15490233 0.15490233 0.17848906
      6      7      8
0.17848906 0.30339158 0.30339158

> n.hyper/n.total
[1] 0.08333333 0.11764706 0.12500000 0.00000000 0.18716578
[6] 0.15294118 0.29411765 0.34782609

> data.frame(fit=fitted(glm.hyp2) * n.total, n.hyper, n.total)
      fit n.hyper n.total
1  5.0267351      5      60
2  1.4242416      2      17
3  1.2392186      1       8
4  0.3098047      0       2
5 33.3774535     35     187
6 15.1715698     13     85
7 15.4729705     15     51
8  6.9780063      8     23
```

There is a large discrepancy for cell 4 between 15% expected (from the model) and 0% observed. However, the expected frequency depends on the number of observations. There are only 2 for cell 4, i.e. that the relative frequency estimate is not reliable. Therefore, it is better to look at counts here.