Name:
DNI/Passport:

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
Departament d'Estadística i Investigació Operativa

MASTER IN DATA SCIENCE (FIB-UPC). ACADEMIC YEAR 23-24 Q1 – FINAL EXAM

Statistical Inference and Modelling (SIM)

Date: 10/Jan/2023 15-18h **Classroom - A6002**

Professor:	Lídia Montero and Josep Franquet
Rules for quiz:	Internet access is not required, emailing and chatting is strictly forbidden. Mobile phones should be switched off. Documents in Final Exam Allowed Document folder on the ATENEA platform can be used.
Duration:	1h 00 min (Part 1) + 2h 30 min (Part 2)
Marks:	Before 22/Jan/24 Subject ATENEA WEB site.
Open Office:	22/Jan/24 at 12:30 – Deganat FIB B6 2nd floor.

Part 1-Problem 1 (10 points): All questions account for the same weight

Suppose x is a single observation on a random variable X ∼ Exp(λ). We wish to test the null hypothesis H0 : 1/λ = 300 against the alternative H1 : 1/λ > 300. We decide to reject H0 if x ≥ 500.

1. What are acceptance/rejection regions A0 and A1?
2. Calculate the probability of Type I error.

Space for H0 is $A_0 = \{0 < x < 500\}$ and the alternative one sided hypothesis $A_1 = \{x \geq 500\}$.

$$P(Type\ I\ Error) = P(Reject\ H0\ |\ H0\ is\ true) = P(\mathrm{x} \geq 500\ |\ \lambda = 0.0033) = \exp(-0.0033 \cdot 500) = 0.18.$$

Distribution function for an exponential distribution with rate paremeter 0.18 is

$$\mathrm{F}(x) = 1 - \exp(-0.0033 \cdot \mathrm{x})$$

Suppose x is a single observation from a random variable X which is distributed X ∼ Binomial(20, π). We wish to test H0 : π = 0.75 against H1 : π < 0.75). We decide to reject H0 if x ≤ 10.

3. What is the acceptance region A0 and the rejection region A1?
4. Calculate the probability of making a Type I error.

Space for H0 is $A_0 = \{10 < x \leq 20\}$ and the alternative one sided hypothesis $A_1 = \{0 < x \leq 10\}$.

$$P(Type\ I\ Error) = P(Reject\ H0\ |\ H0\ is\ true) = P(\mathrm{x} \leq 10\ |\ \pi = 0.75) =$$

$$= \sum_{i=0}^{10} \binom{20}{i} 0.75^i (1-0.75)^{20-i} = 0.103.$$

Probability density function for binomial distribution with paremeters 20 and π=0.75 is

$$\mathrm{p}(i) = \binom{20}{i} 0.75^i (1-0.75)^{20-i}$$

You can use a normal approximation:

$$P(X \leq 10) \approx P\left(Z \leq \frac{10-15+0.5}{\sqrt{20 \cdot 0.75 \cdot 0.25}}\right) = P(Z \leq -1.26) = 0.103$$

Suppose that we have reason to believe that the readings $x_1, x_2, \ldots, x_{16}$ obtained from an experiment were a random sample from a N(μ, σ=3) distribution.

5. We wish to test H0 : μ = μ_0 = 40.0 versus H1 : μ < 40.0. If the observed value of the sample mean is 39.4, what would be the outcome of the test at the 1% significance level?

$$Z = \frac{39.4 - 40}{3/\sqrt{16}} = -0.8$$

The critical value for this one-tailed test at the 1% level is -2.326348 (one-sided test), but -0.8 > -2.33, then we are in the acceptance area.

6. We wish to test H0 : μ = μ_0 = 40.0 versus H1 : μ ≠ 40.0. If the observed value of the sample mean is 44.4, what would be the outcome of the test at the 1% significance level?

$$Z = \frac{44.4 - 40}{3/\sqrt{16}} = 5.86$$

The critical value for this two-tailed test at the 1% level is 2.5758, but 5.86> 2.5758, then we are in the rejection area.

Suppose X ~ N(μ, σ) with σ unknown and let 38.8, 39.2, 39.4, 39.0, 38.6 be a random sample of observations on X.

7. Test at the 1% and 5% level whether μ = 39.5 or not.

Sample mean is $\bar{x}$ =39 and sample variance is 0.1. A two-sided test is proposed and at 5% significance level we reject H0 if $|t|$ > t4(0.975) = 2.776. Thus at 5% H0 is rejected

$$t = \frac{39 - 39.5}{0.3162278/\sqrt{5}} = -3.535$$

At 1% level t4(0.995) = 4.604, thus at 1% level H0 is not rejected too.

We have mild evidence in favour of H0 : μ = 39.5 and against H1 : $\mu \neq 39.5$, because we do reject H0 at the 5% and we do not reject at 1% significance levels. This evidence is mild.

8. Determine a 95% two-sided interval for population variance.

Let us address population variance CI at 95%:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \rightarrow \frac{(5-1)0.1}{\chi^2_{0.025,4}} \leq \sigma^2 \leq \frac{(5-1)0.1}{\chi^2_{0.975,4}} \rightarrow$$

$$\rightarrow \frac{(5-1)0.1}{11.14329} \leq \sigma^2 \leq \frac{(5-1)0.1}{0.4844186} \rightarrow 0.036 \leq \sigma^2 \leq 0.826$$

Name:
DNI/Passport:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

Part 2-Problem 2 (3 points): All questions account for the same weight

Data and models about 50 apartments to be rented in Barcelona (2003) are discussed in this exercise. Price is the target variable.
Variable description:

- Size: in squared meters
- Price: monthly price (euros). Target
- Lift: indicator of lift availability
- Floor: floor in the building. Factor
- Heating: indicator of heating availability
- Air.cond: indicator of air conditioning availability
- Views: indicator whether public space can be seen from any of the windows/balconies of the apartment.
- Bathroom: number of bathrooms
- Furniture: indicator whether is rent including furniture

```
> summary(apartments)
      size            price              lift          floor        rooms          heating
 Min.   : 30.00   Min.   : 600.0   Min.   :0.00   1: 7   Min.   :1.00   Min.   :0.00
 1st Qu.: 56.25   1st Qu.: 727.5   1st Qu.:1.00   2:31   1st Qu.:1.00   1st Qu.:0.00
 Median : 77.50   Median : 850.0   Median :1.00   3:12   Median :2.00   Median :0.00
 Mean   : 76.36   Mean   : 932.4   Mean   :0.82          Mean   :2.24   Mean   :0.48
 3rd Qu.: 95.00   3rd Qu.:1009.4   3rd Qu.:1.00          3rd Qu.:3.00   3rd Qu.:1.00
 Max.   :120.00   Max.   :2350.0   Max.   :1.00          Max.   :5.00   Max.   :1.00
    air.cond        views         bathroom       furniture
 Min.   :0.0   Min.   :0.0   Min.   :1.00   Min.   :0.0
 1st Qu.:0.0   1st Qu.:0.0   1st Qu.:1.00   1st Qu.:0.0
 Median :0.0   Median :1.0   Median :1.00   Median :0.0
 Mean   :0.2   Mean   :0.7   Mean   :1.32   Mean   :0.1
 3rd Qu.:0.0   3rd Qu.:1.0   3rd Qu.:2.00   3rd Qu.:0.0
 Max.   :1.0   Max.   :1.0   Max.   :2.00   Max.   :1.0
```

Two linear models are estimated. The first one includes all the available variables (full model) without including any interactions and the second one is the outcome of applying stepwise regression to the full model.

```
Model A:

Call:
lm(formula = log(price) ~ ., data = apartments)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.108818   0.112717  54.196  < 2e-16 ***
size          0.006143   0.001487   4.132 0.000184 ***
lift          0.061183   0.066296   0.923 0.361742
floor2       -0.030552   0.081272  -0.376 0.709013
floor3        0.173593   0.087732   1.979 0.054945 .
rooms        -0.010854   0.034843  -0.312 0.757058
heating       0.042015   0.064599   0.650 0.519250
air.cond      0.186881   0.065688   2.845 0.007041 **
views        -0.028890   0.056993  -0.507 0.615074
bathroom      0.099189   0.081028   1.224 0.228243
furniture    -0.007238   0.091750  -0.079 0.937526
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1638 on 39 degrees of freedom
Multiple R-squared:  0.7303,     Adjusted R-squared:  0.6612
F-statistic: 10.56 on 10 and 39 DF,  p-value: 2.37e-08
```

```
Model B:
```

```
Call:
lm(formula = log(price) ~ size + floor + air.cond + bathrooms, data = apartments)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.096057   0.090595  67.289  < 2e-16 ***
size         0.006164   0.001290   4.777 2.01e-05 ***
floor2      -0.044450   0.070636  -0.629  0.53242
floor3       0.176584   0.077857   2.268  0.02829 *
air.cond     0.191312   0.058844   3.251  0.00221 **
bathroom     0.131924   0.068591   1.923  0.06092 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1569 on 44 degrees of freedom
Multiple R-squared:  0.7209,     Adjusted R-squared:  0.6892
F-statistic: 22.73 on 5 and 44 DF,  p-value: 3.385e-11
```

Decide and justify whether the next statements are correct, wrong or partially correct:

1. "The best model is model A because R-squared (73.03%) is higher than the one in model B (72.09%)"
2. "In model B, when setting a significance level of 0.1, the variable floor2 is not significant and should be removed from the model"
3. "Since the target variables has been log transformed, then heteroskedasticity has been removed and thus model B can be assumed to have constant variance"

"Since the estimate of air.cond in model B is 0.1913 and the target variable has been log transformed, then it can be interpreted as apartments with air conditioning have a price that is 19.13% greater than one without air conditioning" **Decide and justify whether the next statements are correct, wrong or partially correct:**

1. "The best model is model A because R-squared (73.03%) is higher than the one in model B (72.09%)"

This statement is false. R-Squared is not the only criteria to be accounted for. Redundant variables that are not adding any significantly explicability. B is the output of a stepwise() monitored by AIC, so the lowest AIC corresponds to B and thus it should be preferred.

2. "In model B, when setting a significance level of 0.1, the variable floor2 is not significant and should be removed from the model"

This statement is partially false. Floor factor has 3 levels. Individual pvalues for dummy variables do not have to be taken into account. A binary factor has to be defined grouping 1 and 2 levels and a Fisher test between these 2 models has to be applied to discard current floor factor definition (AIC statistic can also be used to select the model showing lower AIC).

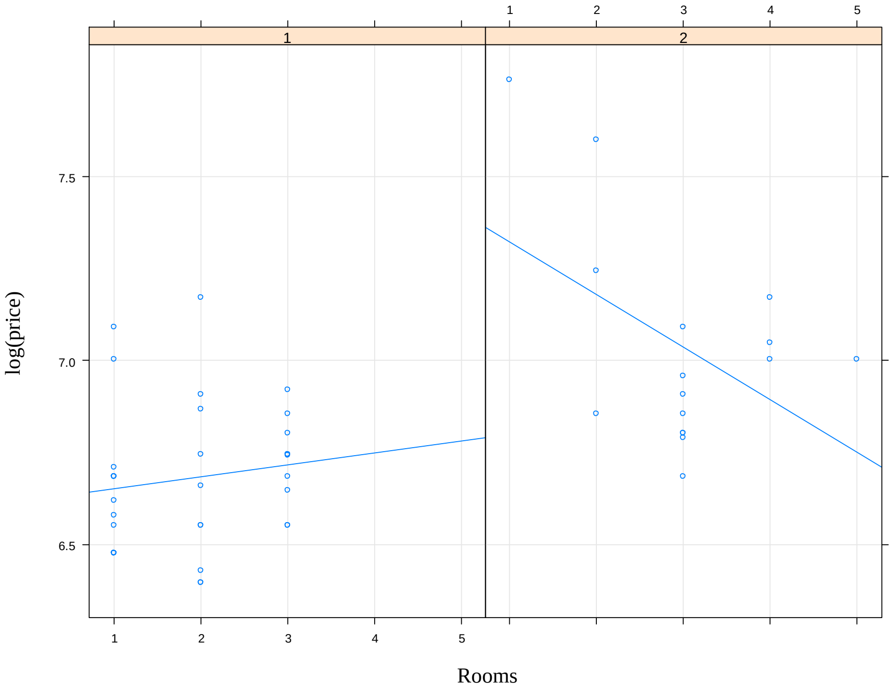
3. "Since the target variables has been log transformed, then heteroskedasticity has been removed and thus model B can be assumed to have constant variance"

False. There is no guarantee that heteroskedaticity has been removed.

4. "Since the estimate of air.cond in model B is 0.1913 and the target variable has been log transformed, then it can *This statement is false.* be interpreted as apartments with air conditioning have a price that is 19.13% greater than one without air conditioning"

This statement is false. The correct answer is exp(0.1913) is 1.210823 and thus with air conditioning the price increases a 21% with respect to no air conditioning apartment, all else being equal. Nevertheless, approximately a percentual interpretation of the air Conditioning parameter may be taken as the increase/decrease in the target scale.

A model for the monthly rental price (log transformed) is estimated based on rooms and bathrooms, taking bathrooms as a categorical variable. Two equations are obtained, the first one considers the relation between price and rooms for apartments with 1 bathroom and the second one does the same for apartment with 2 bathrooms.



The estimated model considering the interaction between rooms and bathroom is the following:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.35160 -0.17205  0.00128  0.10372  0.48627

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.61864    0.09256  71.503  < 2e-16 ***
rooms                0.03261    0.04498   0.725 0.472155
bathroom2            0.84595    0.20529   4.121 0.000156 ***
rooms:bathroom2     -0.17540    0.07364  -2.382 0.021420 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2182 on 46 degrees of freedom
Multiple R-squared:  0.4357,	Adjusted R-squared:  0.3989
F-statistic: 11.84 on 3 and 46 DF,  p-value: 7.165e-06
```

5. Interpret **model equations** and indicate whether the resulting model is reasonable. Predict the monthly price for an apartment of 4 rooms with either 1, or 2 bathrooms.

Model equation for apartment with 1 bathroom:

$$\boldsymbol{log(y) = 6.62 + 0.032\ rooms \rightarrow y = exp(6.62 + 0.032\ rooms) = 749.95\ exp(0.032\ rooms)}$$

Model equation for apartment with 2 bathroom:

$$log(y) = (6.62 + 0.85) + (0.032 - 0.175)\ rooms = 7.47 - 0.143\ rooms$$
$$\rightarrow y = exp(7.47 - 0.143\ rooms) = 1754.61\ exp(-0.143\ rooms)$$

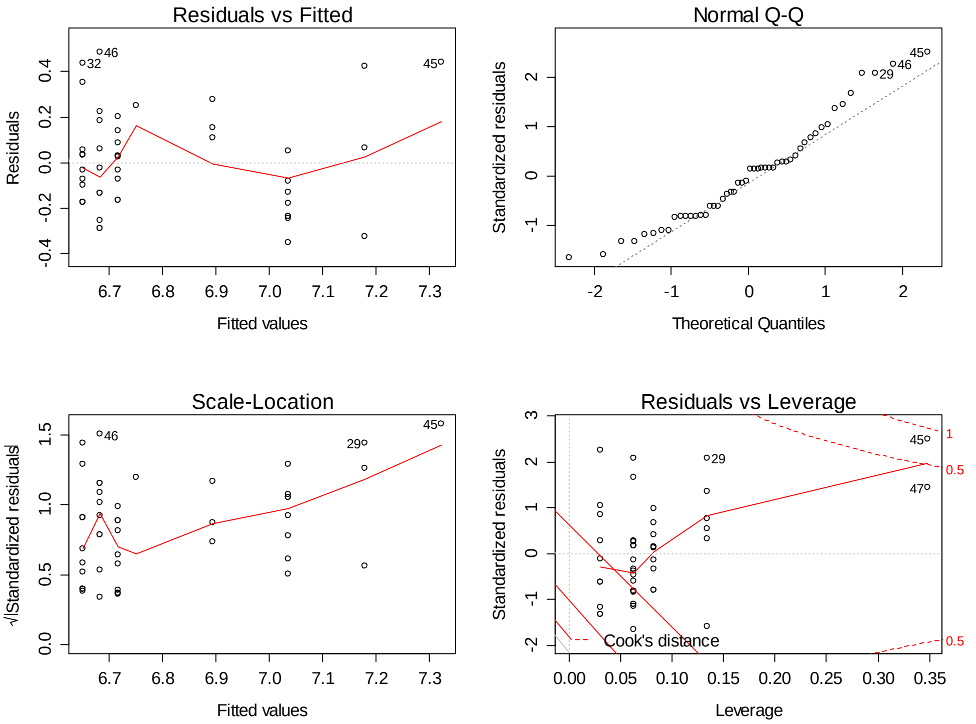
The model is clearly non reasonable, since increasing the number of rooms in 2 bathroom appartments means decreasing the predicted price.

*In the case of a 4 rooms and 1 bathroom, prediction **in the target scale** is:*

$$\hat{y} = exp(6.62 + 0.032 \cdot 4) = exp(6.748) = 852.35€$$

In the case of a 4 rooms and 2 bathroom2

$$\hat{y} = exp(7.47 - 0.143 \cdot 4) = exp(6.898) = 990.29€$$



6. Validate linear model premises based on the available residual analysis plots.

On the top left panel, a pattern in the residual distribution across predicted values can be seen: a transformation of the explanatory variable rooms will be useful. **A random pattern of the residual term is not shown in this model, thus one of the premises is violated.**
On the top right panel, a **deviation from the normal distribution hypothesis** for residuals can be seen in the standardized residuals, at the tails, for small and big values.
According to the scale-location plot, on the left bottom panel, the residual spread (variability) increases as the predicted values increase, thus a **heteroskedastic pattern is seen**.
According to the bottom right panel there are some observations with a large leverage (45 and 47), 45 showing a large residual, and thus both are suspicious of being influent data that have to be removed. Residual outliers have to addressed. **Not a final model, influent data and probably residual outliers can be seen.**

7. Labeled cases in the plots belong to the following observations:

	size	price	lift	floor	rooms	heating	air.cond	views	bathrooms	furniture
29	120	2000	1	3	2	1	1	1	2	0
45	100	2350	1	3	1	1	1	1	2	0
32	95	1200	1	1	1	0	0	0	1	0
47	90	1100	1	2	5	1	0	1	2	0

Indicate for each observation whether it is a residual outlier, or a priori influent data or an actual influent data and detail the effect of each one of them in the model estimation process.

Observation 29 has 2 rooms and 2 bathrooms, heating and air conditioning and the price is expensive (2000€). It is not a residual outlier, nor an influent data.
Observation 45 shows a very expensive apartment with 1 room and 2 bathrooms. It is a residual outlier since the predicted value is not so high and cook's distance is expected to be high since leverage is also high for this observation, so influent data.
Observation 32 is a residual outlier, price is higher than expected according to the model.
Observation 47 has a high leverage and a remarkable positive outlier. It is borderline, but probably influent data.

Name:
DNI/Passport:

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
Departament d'Estadística i Investigació Operativa

Part 2-Problem 3 (4 points): All questions account for the same weight

Wooldridge (2002) analyzes a subset of data collected by Papke in order to assess the impact of investment type on pension plan benefits. The data are available on the Stata website http://www.stata.com/data/jwooldridge/eacsap/pension. There are 226 observations and 21 variables, including some missing, after data cleansing by eliminating the observations with some missing data, 191 observations remain.

Wooldridge Source: L.E.Papke (2004), "Individual Financial Decisionsin Retirement Saving: The Role of Participant-Direction" Journal of Public Economics 88, 39-61. Professor Papke kindly provided the data. She collected them from the National Longitudinal Survey of Mature Women, 1991.

The response variable is polytomous with 3 levels (portfolio typology): "bonds", "mixed" and "stocks" (reference "mixed") and Papke coded these responses based on the percentage included in the discrete quantitative variable **pctstck** for "percentage of publicly traded investment", according to the partition defined by 0, 50 and 100, respectively. In this exercise, the target will be treated in a polytomous way using the variants presented in the course. The choice variable is a dichotomous one that indicates with a 1 whether the person has the possibility to make a choice in the type of investment of the money from their pension fund. Other variables are defined such as age, education, gender, marital status, ethnicity, income, etc. and whether the pension plan is of shared benefits.

```
variable name   type   format      label      variable label
-----------------------------------------------------------------------------
id                        family identifier
years                     years in pension plan
bshared                   =1 if profit sharing plan
choice                    =1 if can choose method invest
female                    =1 if female
married                   =1 if married
age                       age in years
educ                      education years
finc25                    $15,000 < family income 92 <= $25,000
finc35                    $25,000 < family income 92<= $35,000
finc50                    $35,000 < family income 92<= $50,000
finc75                    $50,000 < family income 92<= $75,000
finc100                   $75,000 < family income 92<= $100,000
finc101                   $100,000 < family income 92
wealth89                  assets 1989, $1000
afam                      =1 if afroamerican
stckin89                  =1 if owned stock in 1989
irain89                   =1 if had IRA in 1989
pctstck                   0=mstbnds,50=mixed,100=mststcks
ones                      all ones
target                    c("bonds","mixed","stocks"). Reference category mixed
```

After data cleansing and removal of observations containing NA, a new factor family income is defined in 3 groups <25000, <50000, 50000+ (named f.fincome). Final sample target proportions are 0.3612565 (mixed) 0.3350785 (bonus) 0.3036649 (stocks).

```
> summary(pension[,c(2,3,4,5,6,7,8,15,16,17,18,19,21,23)])
     years         bshared    choice        female      married          age              educ
 Min.   : 0.0    No :151    No : 74    Male  : 75    No : 47     Min.   :54.00    Min.   : 8.00
 1st Qu.: 4.0    Yes: 40    Yes:117    Female:116    Yes:144     1st Qu.:57.00    1st Qu.:12.00
 Median : 9.0                                                    Median :60.00    Median :12.00
 Mean   :11.3                                                    Mean   :60.52    Mean   :13.53
 3rd Qu.:16.0                                                    3rd Qu.:64.00    3rd Qu.:16.00
 Max.   :45.0                                                    Max.   :73.00    Max.   :18.00
    wealth89        afam      stckin89   irain89      pctstck          target        f.fincome
 Min.   :  -6.3   No :169    No :126    No :93    Min.   :  0.00    mixed :69    <=25 mil$:50
 1st Qu.:  65.8   Yes: 22    Yes: 65    Yes:98    1st Qu.:  0.00    bonds :64    <=50 mil$:79
 Median : 140.0                                   Median : 50.00    stocks:58    <50+ mil$:62
 Mean   : 212.0                                   Mean   : 48.43
 3rd Qu.: 253.4                                   3rd Qu.:100.00
 Max.   :1485.0                                   Max.   :100.00
```

Nominal Treatment

1. Determine null model parameter estimates for the polytomous target (mm0). Null deviance is 418.7133 units.

Firstly, data is included in the summary.
Average probability of bonds is 0.3351 (=64/(69+64+58)), odds bonds over mixed are 64/69=0.9275 and logodds log(64/69)= -0.0752.
Average probability of stocks is 0.3036 (=58/(69+64+58)), odds stocks over mixed are 58/69=0.9276 and log(58/69)= -0.1737.

$$(mm0)\ log\left(\frac{\pi_i^b}{\pi_i^m}\right) = \eta_b \rightarrow \widehat{\eta_b} = -0.0752$$

$$log\left(\frac{\pi_i^s}{\pi_i^m}\right) = \eta_s \rightarrow \widehat{\eta_s} = -0.1737$$

The output from R is:

```
> summary(mm0)
Call:
multinom(formula = target ~ 1, data = pension)

Coefficients:
       (Intercept)
bonds  -0.07522289
stocks -0.17366268

Std. Errors:
       (Intercept)
bonds    0.1735447
stocks   0.1781408

Residual Deviance: 418.7133
AIC: 422.7133
```

2. Determine estimated parameters for the multinomial model containing binary factor choice as the explanatory variable (mm1). Residual deviance is 413.153 units.

	target			
choice	**mixed**	**bonds**	**stocks**	
No	21	32	21	***74***
Yes	48	32	37	***117***
	69	***64***	***58***	

$$(mm1)\ \ log\left(\frac{\pi_i^b}{\pi_i^m}\right) = \eta^b + \alpha_i^b \ \ i = 1,2 \ \ i = 1 \equiv \text{choice No} \ \ and \ \ \alpha_1^b = 0$$

$$log\left(\frac{\pi_i^s}{\pi_i^m}\right) = \eta^s + \alpha_i^s \ \ i = 1,2 \ \ i = 1 \equiv \text{choice No} \ \ and \ \ \alpha_1^s = 0$$

For bonds equation: $i = 1 \equiv \text{No} \ \ \hat{\eta}^b \ = log\frac{32}{21} = 0.4212$

$i = 2 \equiv \text{Yes} \ \ \hat{\alpha}_2^b = \hat{\eta}^b \ + \hat{\alpha}_2^b - \hat{\eta}^b \ = log\frac{32}{48} - log\frac{32}{21} = -0.8267$

For stocks equation: $i = 1 \equiv \text{choice No} \ \ \hat{\eta}^s \ = log\frac{21}{21} = 0$

$i = 2 \equiv \text{choice Yes} \ \ \hat{\alpha}_2^s = \hat{\eta}^s \ + \hat{\alpha}_2^s - \hat{\eta}^s \ = log\frac{37}{48} - log\frac{21}{21} = -0.2603$

```
> summary(mm1)
Call:
multinom(formula = target ~ choice, data = pension)

Coefficients:
         (Intercept)   choiceYes
bonds   4.212139e-01 -0.8266792
```

```
stocks 6.980111e-07 -0.2602834

Std. Errors:
        (Intercept) choiceYes
bonds     0.2808364 0.3618735
stocks    0.3086067 0.3782836

Residual Deviance: 413.153
AIC: 421.153
```

3. Address a deviance test to determine whether choice factor is significant or not in the target level proportions.

A deviance test stating H0:'Models (mm0) and (mm1) are equivalent' based of the asymptotic distribution of $\Delta Dev(mm0, mm1)$ as $\chi^2_{v=2}$ has to be addressed.

Residual deviance for the null model (mm0) is 418.71 and residual deviance for (mm1) is 413.15 according to provided data:

$$\Delta Dev(mm0, mm2) = Dev(mm0) - Dev(mm1) = 418.7133 - 413.153 = 5.560296$$

Thus $P(\chi^2_2 > 5.560296) = 0.0620$ and **H0 can not be rejected at the 0.05 significance level, probabilities of target categories do not depend on choice (too close to the significance level)**

The results of fitting the additive multinomial logit model (mm2) using choice, bshared, wealth89, age, educ, female, married, afam and f.fincome factors are presented below. It has a pseudo-coefficient of determination (McFadden) of 0.087. By removing the main effect of choice, the logarithm of the likelihood goes down by 2.2 units:

bonus vs mixed	Estimates	stocks vs mixed	Estimates
(Intercept)	-1.3732	**(Intercept)**	2.29456
choiceYes	-0.68088	**choiceYes**	0.098152
bsharedYes	0.27954	**bsharedYes**	1.216165
wealth89	0.000643	**wealth89**	0.000353
age	0.080538	**age**	-0.01161
educ	-0.12711	**educ**	-0.07875
femaleFemale	-0.33613	**femaleFemale**	-0.14733
marriedYes	-0.73279	**marriedYes**	-0.44459
afamYes	-0.64059	**afamYes**	-0.16961
f.fincome<=50	-1.16233	**f.fincome<=50**	-0.62821
f.fincome>50+	-0.94837	**f.fincome>50+**	-1.02328
LogLik	**-191.144**	**LogLik Null Model**	**-209.3567**
Explained Deviance	**36.42524**	**Residual Deviance**	

4. Formally state the model. Detail the number of parameters of the additive model. What is the residual deviance of the additive model mm2?

There are 2 logodd equations 1) Bonds vs Mixed 2) Stocks vs Mixed.
For each equation the number of parameters are 11, thus 11x2=22 parameters.
Residual deviance is twice minus the logLik function value Dev(mm2)=2*(-logLik(mm2))=2*(191.144)= 382.2881.

(mm2) (Bonds vs Mixed) $\log\left(\frac{\pi^b_{ijklmn}}{\pi^m_{ijklmn}}\right) = \eta + \alpha_i + \beta_j + \gamma_k + \delta_l + \varepsilon_m + \rho_n + \kappa \cdot wealth89 + \mu \cdot age + \nu \cdot educ$

(Stocks vs Mixed) $\log\left(\frac{\pi^s_{ijklmn}}{\pi^m_{ijklmn}}\right) = \eta' + \alpha'_i + \beta'_j + \gamma'_k + \delta'_l + \varepsilon'_m + \rho'_n + \kappa' \cdot wealth89 + \mu' \cdot age + \nu' \cdot educ$

Where $\alpha_1 = 0$ and α_2 for choice Yes
Where $\beta_1 = 0$ and β_2 for bshared Yes
Where $\gamma_1 = 0$ and γ_2 for female yes
Where $\delta_1 = 0$ and δ_2 for married Yes
Where $\varepsilon_1 = 0$ and ε_2 for afam Yes
Where $\rho_1 = 0$ and ρ_2 for f.fincome <=50 and ρ_3 for f.fincome >50+

And the same dummy variable statement applies for prime (') variables in the second log odd equation (stock versus mixed).

5. Interpret the effect of choice on the outcome in terms of logodds and relative probabilities (odds).

- In the case of logodd equation for bonds vs mixed choice-yes adds -0.68088 units compared to **choice-No-reference level, all else being equal.**
- In the case of logodd equation for stocks vs mixed choice-yes adds 0.098152 units compared to **choice-No-reference level, all else being equal.**
- In the case of odds for bonds vs mixed choice-yes the effect is multiplicative by 0.5061691 =exp(-0.68088) **with respect to choice-No-**reference level, **all else being equal.** So, relative probability of bonds vs mixed decreases by 49.39% **with respect to choice-no all else being equal.**
- In the case of odds for stocks vs mixed choice-yes the effect is multiplicative by 1.1031300 =exp(0.098152) **with respect to choice-No-**reference level, **all else being equal.** So, relative probability of bonds vs mixed increases by 10.31% **with respect to choice-no all else being equal.**

6. What are the predicted probabilities for the response categories for an afro-american unmarried man having an annual income over 50000$ without shared benefit (bshared), nor choice in the mean for the numeric variables in mm2?

i=1 (choice No), j=1 (bshared No), k=1 (man), l=1(unmarried), m=2 (afam Yes) and 3 (f.fincome >50+) refer to index meaning in model statement. Mean values for covariates are for wealth89, age and educ 212.0, 60.52 and 13.53 respectively.

(mm2) $$\log\left(\frac{\pi^b_{ijklmn}}{\pi^m_{ijklmn}}\right) = \eta + \alpha_i + \beta_j + \gamma_k + \delta_l + \varepsilon_m + \rho_n + \kappa \cdot wealth89 + \mu \cdot age + \nu \cdot educ$$
$$\log\left(\frac{\pi^s_{ijklmn}}{\pi^m_{ijklmn}}\right) = \eta' + \alpha'_i + \beta'_j + \gamma'_k + \delta'_l + \varepsilon'_m + \rho'_n + \kappa' \cdot wealth89 + \mu' \cdot age + \nu' \cdot educ$$

$$\log\left(\frac{\pi^b_{111123}}{\pi^m_{111123}}\right) = \eta + \alpha_1 + \beta_1 + \gamma_1 + \delta_1 + \varepsilon_2 + \rho_3 + \kappa \cdot 212 + \mu \cdot 60.52 + \nu \cdot 13.53 =$$
$$= -1.3732 + 0 + 0 + 0 + 0 - 0.64059 - 0.94837 + 0.000643 \cdot 212 + 0.080538 \cdot 60.52 - 0.12711 \cdot 13.53 = 0.3285$$
$$\rightarrow \frac{\pi^b_{111123}}{\pi^m_{111123}} = \exp(0.3285) = 1.3888$$

$$\log\left(\frac{\pi^s_{111123}}{\pi^m_{111123}}\right) = \eta' + \alpha'_i + \beta'_j + \gamma'_k + \delta'_l + \varepsilon'_m + \rho'_n + \kappa' \cdot wealth89 + \mu' \cdot age + \nu' \cdot educ =$$
$$= 2.294560 + 0 + 0 + 0 + 0 - 0.1696098 - 1.0232754 + 0.0003531203 * 212 - 0.01160917 * 60.52 - 0.07875259 * 13.53 =$$
$$= -0.5915732$$
$$\rightarrow \frac{\pi^s_{111123}}{\pi^m_{111123}} = \exp(-0.5915732) = 0.5534559$$

$$\pi^m_{111123} = \frac{1}{1 + \frac{\pi^b_{111123}}{\pi^m_{111123}} + \frac{\pi^s_{111123}}{\pi^m_{111123}}} = 0.3398729$$
$$\pi^b_{111123} = \pi^m_{111123} \frac{\pi^b_{111123}}{\pi^m_{111123}} = 0.4720224$$
$$\pi^s_{111123} = \pi^m_{111123} \frac{\pi^s_{111123}}{\pi^m_{111123}} = 0.1881047$$

```
> predict(mm2,type="probs",newdata=data.frame(choice="No",bshared="No",female="Ma
le",married="No",afam="Yes",f.fincome=">50+ mil$",wealth89=212,age=60.52,educ=13.
53))
    mixed     bonds    stocks
0.3398729 0.4720224 0.1881047
```

Name:

DNI/Passport:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

Ordinal Treatment (om2 proportional odds model)

Coefficients	Estimates (latent)	Standard error
choiceYes	0.084643	0.296594
bsharedYes	0.948004	0.351477
wealth89	0.000325	0.000595
age	-0.00633	0.0367
educ	-0.05784	0.056557
femaleFemale	-0.15311	0.340035
marriedYes	-0.30775	0.37835
afamYes	-0.13857	0.469284
f.fincome<=50	-0.40325	0.355781
f.fincome<50+	-0.75833	0.44334
Constant mixed|bonds	**-2.1951**	2.5098
Constant bonds|stocks	**-0.6965**	2.5051
LogLik	-201.2246	
LogLik Null Model	-209.3567	
Residual Deviance	402.4491	
Null Deviance	418.7133	

7. Formulate the model. Detail the number of parameters of the additive model. Use level order as mixed, bonus and stocks in all the sections.

(om2) $$log\left(\frac{\gamma^m_{ijklmn}}{1-\gamma^m_{ijklmn}}\right) = \eta^m + \alpha_i + \beta_j + \gamma_k + \delta_l + \varepsilon_m + \rho_n + \kappa \cdot wealth89 + \mu \cdot age + \nu \cdot educ$$

$$log\left(\frac{\gamma^b_{ijklmn}}{1-\gamma^b_{ijklmn}}\right) = \eta^b + \alpha_i + \beta_j + \gamma_k + \delta_l + \varepsilon_m + \rho_n + \kappa \cdot wealth89 + \mu \cdot age + \nu \cdot educ$$

Additive model number of parameters is p=2+10=12. Supraindex m refers to mixed category and b to bonds category of the target variable.
Where $\alpha_1 = 0$ and α_2 for choice Yes
Where $\beta_1 = 0$ and β_2 for bshared Yes
Where $\gamma_1 = 0$ and γ_2 for female yes
Where $\delta_1 = 0$ and δ_2 for married Yes
Where $\varepsilon_1 = 0$ and ε_2 for afam Yes
Where $\rho_1 = 0$ and ρ_2 for f.fincome <=50 and ρ_3 for f.fincome >50+

8. Interpret the effect of choice in terms of proportional odds and latent variable.

In terms of latent variable analysis, coefficients are directly the ones provided and propension to stocks increases in the logodds scale for choice-yes by 0.084643 units. If we divide these estimates by the standard deviation of standard logistic scale $\sqrt{\pi^2/3}$ =1.814, then we have the effect on standard deviation times in the logistic scale assumed for the propension variable. So, choice-Yes has an effect on the propension scale of moving mixed or bonds 0.0466 standard deviations to the right in the propension scale with respect to the reference category choice-No all else being equol.

Under the proportional odds point of view, choice-Yes decreases by 8.116% odds of mixed vs (bonds or stocks) and the odds of mixed or bonds versus stocks, all else being equal.

```
> exp(-0.084642808)
[1] 0.9188404
> (1-exp(-0.084642808))*100
[1] 8.115957
```

9. What are the predicted probabilities for the response categories for an afro-american unmarried man having an annual income over 50000$ without shared benefit (bshared), nor choice in the mean for the numeric variables based on om2?

i=1 (choice No), j=1 (bshared No), k=1 (man), l=1(unmarried), m=2 (afam Yes) and 3 (f.fincome >50+) refer to index meaning in model statement. Mean values for covariates are for wealth89, age and educ 212.0, 60.52 and 13.53 respectively,

$$log\left(\frac{\gamma^m_{111123}}{1-\gamma^m_{111123}}\right) = log\left(\frac{\pi^m_{111123}}{\pi^b_{111123}+\pi^s_{111123}}\right) =$$
$$= -2.1951 + 0 + 0 + 0 + 0 + 0.138568224 + 0.758329730 - 0.000325147 * 212 + 0.006332663 * 60.52 + 0.057841010 * 13.53 = -0.2012916$$
$$\rightarrow \gamma^m_{111123} = \frac{exp(-0.2012916)}{1+exp(-0.2012916)} = 0.4498463$$

$$log\left(\frac{\gamma^b_{111123}}{1-\gamma^b_{111123}}\right) = log\left(\frac{\pi^m_{111123}+\pi^b_{111123}}{\pi^s_{111123}}\right) = -0.6965 + 0 + 0 + 0 + 0 + 0.138568224 + 0.758329730 - 0.000325147 * 212 + 0.006332663 * 60.52 + 0.057841010 * 13.53 = 1.297308 \rightarrow \gamma^b_{111123} = \frac{exp(1.297308)}{1+exp(1.297308)} = 0.7853816 \rightarrow \pi^b_{111123} = \gamma^b_{111123} - \gamma^m_{111123} = 0.7853816 - 0.4498463 = 0.3355353$$
$$\pi^s_{111123} = 1 - \gamma^b_{111123} = 1 - 0.7853816 = 0.2146184$$

So, mixed probabilily 0.450, bonds probability 0.335 and stocks probability 0.215

10. Compare the nominal/ordinal additive proposals according to Akaike's criterion.

Mininum AIC is obtained by the nominal proposal (marginally).
AIC(nominal)=2*(-logLik(nominal) + p(nominal))= 2(191.144+22)= 426.2881
AIC(ordinal)= 2*(-logLik(ordinal) + p(ordinal))= 2(201.2246+12)= 426.4491

Name:

DNI/Passport:

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
Departament d'Estadística i Investigació Operativa

Part 2-Problem 4 (3 points): All questions account for the same weight

The Insurance data set in MASS library contains the number of claims between customers (policies) of a British car insurance company in 1973. The description of the columns is as follows:

`District`	district of policyholder (1 to 4): 4 is major cities (London).
`Group`	group of car (1 to 4), <1 litre, 1–1.5 litre, 1.5–2 litre, >2 litre.
`Age`	of driver in 4 ordered groups, <25, 25–29, 30–35, >35.
`Holders`	numbers of policyholders (pòlisses)
`Claims`	numbers of claims (*sinistres*)

Source: L. A. Baxter, S. M. Coutts and G. A. F. Ross (1980) Applications of linear models in motor insurance. *Proceedings of the 21st International Congress of Actuaries, Zurich* pp. 11–29

```
> summary(Insurance)
 District     Group        Age         Holders           Claims
 1:16     <1l    :16   <25  :16   Min.   :   3.00   Min.   :  0.00
 2:16     1-1.5l:16    25-29:16   1st Qu.:  46.75   1st Qu.:  9.50
 3:16     1.5-2l:16    30-35:16   Median : 136.00   Median : 22.00
 4:16     >2l    :16   >35  :16   Mean   : 364.98   Mean   : 49.23
                                  3rd Qu.: 327.50   3rd Qu.: 55.50
                                  Max.   :3582.00   Max.   :400.00
```

The data corresponds to 23359 policy holders where 3151 claims have been reported. The authors indicated as source, analyze the data using loglinear models with the number of claims as response and the number of policies as offset. You have some results from R below.

1. Assess the net effects of the available factors in the additive Poisson model and statistically justify whether it is possible to delete any term in the model: Claims ~ offset (logtamany) + District + Age + Group.

According to the provided output, all net-effects are significant (District, Group and Age), so it is not possible to remove any explanatory factor.

```
> Anova(ma)
Analysis of Deviance Table (Type II tests)

Response: Claims
         LR Chisq Df Pr(>Chisq)
District   13.871  3   0.003086 **
Group      88.667  3  < 2.2e-16 ***
Age        84.870  3  < 2.2e-16 ***
```

2. Apply a goodness of fit test to the Poisson additive model.

Deviance for the additive model is 51.42 and 54 df, distributed as Chisq Distribution with 54 df.

```
> # GoF: H0 Model is consistent to data  H0  can not be Rejected (Accepted)
> # Residual Deviance test statistic
> 1-pchisq(ma$deviance,ma$df.residual)
[1] 0.5745071
```

3. Predict the expected number of claims for a London policy holder in the youngest age group and car group <1 litre. What is the probability of reporting one o more claims in 1973 for such a policy holder?

$$log(y_{ijk}) = \eta + \alpha_i + \beta_j + \gamma_k \qquad \begin{array}{c} \alpha_1 = 0 \; for \; District \; 1 \\ \beta_1 = 0 \; for \; Group < 1l \\ \gamma_1 = 0 \; for \; Age \; group \; < 25 \end{array}$$

$$log(y_{411}) = \eta + \alpha_4 + \beta_1 + \gamma_1 = -1.82173992 + 0.23420533 + 0 + 0 = -1.587535$$

$\rightarrow \hat{y}_{411} = \exp(-1.587535) = 0.204429$

The expected number of claims for the selected individual is 0.204.
The probability for a Poisson Distribution to report one or more claims in 1973 can be obtained as 1 minus the complementary succés (0 claims)

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - \frac{(0.204429)^0}{0!}\exp(-0.204429) = 0.1848874$$

4. The estimated model supports a Poisson response. How could you validate this hypothesis? Would the conclusions change much in the presence of overdispersion? Estimate the overdispersion parameter.

Generalized Pearson statistics for the additive model is 48.62934 and assympotically distributed as a Chisquared with 54 df. Overdispersion parameter estimate takes the value 48.62934/54= 0.9. A dispersion test to determine the consistency to the negative binomial distribution can not be proved to hold. Null hypothesis is alpha=0 can not be rejected.

$$h(\mu_i) = \mu_i^2 \rightarrow V[Y_i | X_i] = \mu_i + \alpha\mu_i^2 = (1 + \alpha\mu_i)\mu_i$$

```
> dispersiontest(ma,trafo=2)
         Overdispersion test data:  ma
z = -1.8988, p-value = 0.9712
alternative hypothesis: true alpha is greater than 0
```

5. The negative binomial response additive model is included. Test whether the additive model does the same job as the null model filling the table of the deviance test given below.

```
Analysis of Deviance Table
Model 1: Claims ~ offset(logsize)
Model 2: Claims ~ offset(logsize) + District + Group + Age
```

	Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
1	63	(1)				
2	54	(2)	(3)	(4)	(5)	(6)

```
> anova(mabn0,mabn,test="F")
Analysis of Deviance Table

Model 1: Claims ~ offset(logsize)
Model 2: Claims ~ offset(logsize) + District + Group + Age
  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1        63    236.212
2        54     51.416  9    184.8 22.802 2.137e-15 ***
```

Null hypothesis is rejected, thus the additive model is substantially better to the null model.

6. Predict the expected number of claims for a London policy holder in the youngest age group and car group <1 litre according to the negative binomial additive model.

$$log(y_{ijk}) = \eta + \alpha_i + \beta_j + \gamma_k \qquad \begin{matrix} \alpha_1 = 0 \; for \; District \; 1 \\ \beta_1 = 0 \; for \; Group < 1l \\ \gamma_1 = 0 \; for \; Age \; group \; < 25 \end{matrix}$$

$$log(y_{411}) = \eta + \alpha_4 + \beta_1 + \gamma_1 = -1.82173983 + 0.23420601 + 0 + 0 = -1.587534$$

$\rightarrow \hat{y}_{411} = \exp(-1.587534) = 0.2044291$

The expected number of claims for the selected individual is 0.204.

7. The gamma response additive model is included. Test whether the additive model does the same job as the null model filling the table of the deviance test given below.

```
Analysis of Deviance Table
```

```
Model 1: Claims ~ offset(logsize)
Model 2: Claims ~ offset(logsize) + District + Group + Age

  Resid. Df Resid. Dev     Df     Deviance      F       Pr(>F)
1         63     (1)[    ]
2         54     (2)[    ]    (3)[ ]    (4)[    ] (5)[    ] (6)[    ]
```

```
> anova(baxter.gma0,baxter.gma,test="F")
Analysis of Deviance Table

Model 1: I(Claims + 0.5) ~ offset(logsize)
Model 2: I(Claims + 0.5) ~ offset(logsize) + District + Age + Group
  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1        63    10.7539
2        54     3.9161  9   6.8378 12.344 2.427e-10 ***
```

Null hypothesis is rejected, thus the additive model is substantially better to the null model.

8. Is the District factor net effect significant in the additive gamma model at the 5% significance level?

Taking the results below about the net-effect tests on baxter.gma (gamma additive model), at the 5% significance level District factor net-effect is not significant.

```
> Anova(baxter.gma,test="F")
Analysis of Deviance Table (Type II tests)

Response: I(Claims + 0.5)
Error estimate based on Pearson residuals

          Sum Sq Df F value    Pr(>F)
District  0.4267  3  2.3109   0.08649 .
Age       2.6495  3 14.3491 5.409e-07 ***
Group     3.7194  3 20.1441 6.856e-09 ***
Residuals 3.3236 54
```

9. Predict the expected number of claims for a London policy holder in the youngest age group and car group <1 litre according to the gamma additive model.

$$log(y_{ijk}) = \eta + \alpha_i + \beta_j + \gamma_k \qquad \begin{matrix} \alpha_1 = 0 \; for \; District \; 1 \\ \beta_1 = 0 \; for \; Group < 1l \\ \gamma_1 = 0 \; for \; Age \; group \; < 25 \end{matrix}$$

$$log(y_{411}) = \eta + \alpha_4 + \beta_1 + \gamma_1 = -1.8382672 \; + \; 0.2292878 \; + 0 + 0 = -1.608979$$

$$\rightarrow \; \hat{y}_{411} = \exp(-1.608979) = 0.2000$$

The expected number of claims for the selected individual is 0.200.

Results

```
> ma<-glm(Claims~offset(logsize)+District+Group+Age, family=poisson,data=df)
> summary(ma)

Call: glm(formula = Claims ~ offset(logsize) + District + Group + Age, family = poisson, data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.82174    0.07679 -23.724  < 2e-16 ***
District2    0.02587    0.04302   0.601 0.547597
District3    0.03852    0.05051   0.763 0.445657
District4    0.23421    0.06167   3.798 0.000146 ***
Group1-1.5l  0.16134    0.05053   3.193 0.001409 **
Group1.5-2l  0.39281    0.05500   7.142 9.18e-13 ***
Group>2l     0.56341    0.07232   7.791 6.65e-15 ***
Age25-29    -0.19101    0.08286  -2.305 0.021149 *
Age30-35    -0.34495    0.08137  -4.239 2.24e-05 ***
Age>35      -0.53667    0.06996  -7.672 1.70e-14 ***
---

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

```
    Null deviance: 236.26  on 63  degrees of freedom
Residual deviance:  51.42  on 54  degrees of freedom
AIC: 388.74

> Anova(ma)
Analysis of Deviance Table (Type II tests)

Response: Claims
         LR Chisq Df Pr(>Chisq)
District   13.871  3   0.003086 **
Group      88.667  3  < 2.2e-16 ***
Age        84.870  3  < 2.2e-16 ***

> X2P<-sum(resid(ma,type="pearson")^2);X2P
[1] 48.62934

> dispersiontest(ma,trafo=2)
         Overdispersion test data:  ma
z = -1.8988, p-value = 0.9712
alternative hypothesis: true alpha is greater than 0


> mabn<-glm(Claims~offset(logsize)+District+Group+Age,family=neg.bin(449934),data=df)
> summary(mabn)

Call:
glm(formula = Claims ~ offset(logsize) + District + Group + Age,     family = neg.bin(449934), data =
df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.82174    0.07287 -25.000  < 2e-16 ***
District2    0.02587    0.04083   0.634 0.528976
District3    0.03853    0.04794   0.804 0.425114
District4    0.23421    0.05853   4.002 0.000193 ***
Group1-1.5l  0.16133    0.04796   3.364 0.001419 **
Group1.5-2l  0.39281    0.05219   7.526 5.77e-10 ***
Group>2l     0.56341    0.06863   8.210 4.53e-11 ***
Age25-29    -0.19101    0.07863  -2.429 0.018487 *
Age30-35    -0.34495    0.07722  -4.467 4.09e-05 ***
Age>35      -0.53667    0.06639  -8.084 7.22e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial family taken to be [         ])

    Null deviance: 236.212  on 63  degrees of freedom
Residual deviance:  51.416  on 54  degrees of freedom
AIC: 388.74

Number of Fisher Scoring iterations: 4

> Anova(mabn, test="F")
Analysis of Deviance Table (Type II tests)
Response: Claims
Error estimate based on Pearson residuals

          Sum Sq Df F value    Pr(>F)
District  13.869  3   5.134  0.003387 **
Group     88.651  3  32.816 3.263e-12 ***
Age       84.856  3  31.412 6.904e-12 ***
Residuals 48.626 54
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> AIC(mabn, ma)
     df      AIC
mabn 10 388.7450
ma   10 388.7416
> baxter.gma0 <- glm(I(Claims+0.5)~offset(logsize), family=Gamma(link=log),data=df)
> baxter.gma <- glm(I(Claims+0.5)~offset(logsize)+District+Age+Group, family=Gamma(link=log),data=df)
> summary(baxter.gma)

Call:
glm(formula = I(Claims + 0.5) ~ offset(logsize) + District +
    Age + Group, family = Gamma(link = log), data = df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.83827    0.09807 -18.745  < 2e-16 ***
District2    0.14271    0.08771   1.627 0.109553
District3    0.11118    0.08771   1.268 0.210413
District4    0.22929    0.08771   2.614 0.011569 *
Age25-29    -0.22686    0.08771  -2.586 0.012427 *
Age30-35    -0.36304    0.08771  -4.139 0.000123 ***
Age>35      -0.56083    0.08771  -6.394 3.96e-08 ***
Group1-1.5l  0.13818    0.08771   1.575 0.121023
Group1.5-2l  0.42257    0.08771   4.818 1.22e-05 ***
Group>2l     0.61872    0.08771   7.054 3.37e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.06154746)
```

```
    Null deviance: 10.7539  on 63  degrees of freedom
Residual deviance:  3.9161  on 54  degrees of freedom
AIC: 430.38

Number of Fisher Scoring iterations: 5

> Anova(baxter.gma,test="F")
Analysis of Deviance Table (Type II tests)

Response: I(Claims + 0.5)
Error estimate based on Pearson residuals

          Sum Sq Df F value    Pr(>F)
District  0.4267  3  2.3109   0.08649 .
Age       2.6495  3 14.3491 5.409e-07 ***
Group     3.7194  3 20.1441 6.856e-09 ***
Residuals 3.3236 54
```