Name:
DNI/Passport:

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

MASTER IN DATA SCIENCE (FIB-UPC). ACADEMIC YEAR 22-23 Q1 – FINAL EXAM

Statistical Inference and Modelling (SIM) .

Date: 11/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 20/Jan/23 Subject ATENEA WEB site.
Open Office:	23/Jan/23 – 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 = 100 against the alternative H1 : 1/λ > 100. We decide to reject H0 if x ≥ 460.

1. What are the acceptance region A0 and the rejection region A1?
2. Calculate the probability of making a Type I error.

Space for HO is $A_0 = \{0 < x < 460\}$ *and the alternative one sided hypothesis* $A_1 = \{x \geq 460\}$.

$$P(Type\ I\ Error) = P(Reject\ H0 \mid H0\ is\ true) = P(x \geq 460 \mid \lambda = 0.01) = \exp(-0.01 \cdot 460) = 0.01.$$

Distribution function for an exponential distribution with rate paremeter 0.01 is

$$F(x) = 1 - \exp(-0.01 \cdot x)$$

Suppose x is a single observation from a random variable X which is distributed X ~ Binomial(20, π). We wish to test H0 : π = 0.5 against H1 : π < 0.5). We decide to reject H0 if x ≤ 6.

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

Space for HO is $A_0 = \{6 < x \leq 20\}$ *and the alternative one sided hypothesis* $A_1 = \{0 < x \leq 6\}$.

$$P(Type\ I\ Error) = P(Reject\ H0 \mid H0\ is\ true) = P(x \leq 6 \mid \pi = 0.5) =$$

$$= \sum_{i=0}^{6} \binom{20}{i} 0.5^i (1-0.5)^{20-i} = 0.0577.$$

Probability density function for binomial distribution with paremeters 20 and π=0.5 is

$$p(i) = \binom{20}{i} 0.5^i (1-0.5)^{20-i}$$

You can use a normal approximation:

$$P(X \leq 6) \approx P\left(Z \leq \frac{6-10+0.5}{\sqrt{20 \cdot 0.5 \cdot 0.5}}\right) = P(Z \leq -1.56) = 0.059$$

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=2) distribution.

5. We wish to test H0 : μ = μ_0 = 36.0 versus H1 : μ < 36.0. If the observed value of the sample mean is 34.4, what would be the outcome of the test at the 5% significance level?

$$Z = \frac{34.4 - 36}{2/\sqrt{16}} = -3.2$$

The critical value for this one-tailed test at the 5% level is -1.645 (one-sided test), but -3.2 < 1.645, then we are in the rejection area.

6. We wish to test H0 : μ = μ_0 = 36.0 versus H1 : μ ≠ 36.0. If the observed value of the sample mean is 34.4, what would be the outcome of the test at the 1% significance level?

$$Z = \frac{34.4 - 36}{2/\sqrt{16}} = -3.2$$

The critical value for this two-tailed test at the 1% level is -2.5758, but -3.2 < -2.5758, then we are in the rejection area again.

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 = 40 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

$$t = \frac{39 - 40}{0.01/\sqrt{5}} = -7.07$$

At 1% level $t4(0.995) = 4.604$, thus at 1% level H0 is also rejected.

At 0.1% level $t4(0.9995) = 8.61$, thus at 0.1% level H0 can not be rejected.

We have evidence against H0 : $\mu = 40$ in favour of H1 : $\mu \neq 40$, because we reject H0 at the 5% level. This evidence is strong, because we reject H0 at the 1% level, but not overwhelming, because we do not reject H0 at the 0.1% level.

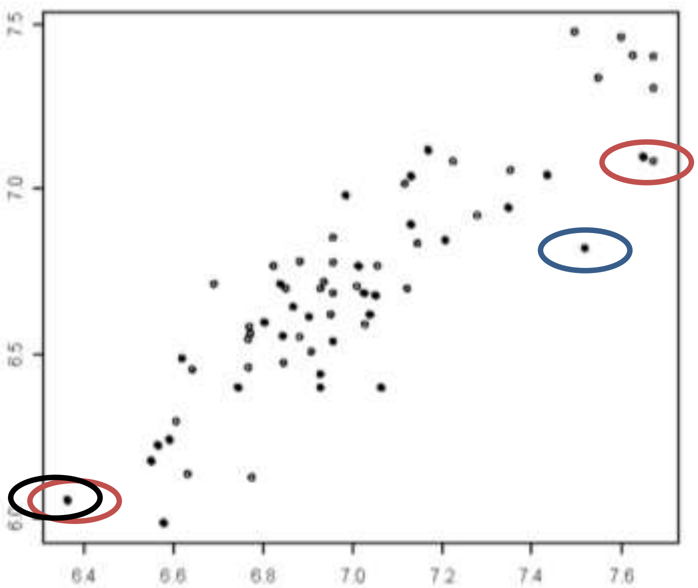
8. Determine a 95% two-sided interval for population variance.

Let us address population variance CI at 95%:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \rightarrow \frac{(5-1)0.1^2}{\chi^2_{0.025,4}} \leq \sigma^2 \leq \frac{(5-1)0.1^2}{\chi^2_{0.975,4}} \rightarrow$$
$$\rightarrow \frac{(5-1)0.1^2}{11.14329} \leq \sigma^2 \leq \frac{(5-1)0.1^2}{0.4844186} \rightarrow 0.0036 \leq \sigma^2 \leq 0.0826$$

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

Taxes to be paid on the purchase of luxury vehicles depend on their market price. Data are available on 66 luxury vehicles sold second-hand in Barcelona. The linear relationship between the logarithm of the tax amount (log_rate, in thousands of euros), as the response variable, and the logarithm of the price (log_price, in thousands of euros), as an explanatory variable, graphically looks like satisfactory. Additionally, there is a dichotomous variable that indicates whether the vehicle is less than 3 years old (new) or not (old). A baseline re-parameterization is proposed for Age factor where the reference group is vehicles with 3 or more years (old) of age (indicated with black dots).



Blue – Residual outlier
Black – High leverage
Brown– Influential data

Below are detailed technical aspects of linear models:

• The mean of log_rate is 6.7128 and the standard deviation 0.3410.

• The simple linear regression model has a slope of 0.8161. If the two variables are standardized (after centered on their mean), the slope becomes 0.8824 (which illustrates the linear correlation coefficient directly). The residual sum of squares is 1.6732.
• The general linear model with only Age factor provides a coefficient of 0.1667 for the dummy variable with a standard error of 0.0826.
• The additive Ancova model has a residual sum of squares of 1.5692 and equation: log_rate = 0.1080+0.9355 log_price + 0.0806 Dummy_New .
• The complete Ancova model reduces the residual sum of squares by 0.0207 units.

Answer the following questions while justifying them statistically based on the topics that have been seen in the course.

1. Interpret the slope in the simple linear regression model.

A 1% increase in price is associated with a 0.8% increase in taxes. Given that the increase is less than 1%, the vehicle tax is regressive, not progressive, meaning that more expensive cars pay proportionally less tax.
rate = exp(b1)*exp(0.8161*log_price)= exp(b1)*(exp(log_price))^ 0.8161=
exp(b1)*(price)^0.8161.
And hence an increase in the price of 1% implies an increase in the rate of (1.01)^0.8161= 1.00815, that is, an increase of 0.8%.
In the log scale, an increament of 1 unit in log_price represents an increment of 0.8161 units in log_tax

2. In view of the information do you think that new vehicles pay more taxes than old vehicles of the same price? Interpret the coefficient of the dummy variable in the additive Ancova model and test its statistical significance.

The coefficient of 0.0806 associated to new vehicles **in the additive Ancova model** (general linear model can not be taken since it contains gross-effect only) is positive and therefore indicates an multiplicative effect of 8% on the rates to be paid (additive effect of 0.0806 log units), for new cars in the same price, which seems negligible. It must be formally contrasted by Fisher test. The null hypothesis is that there is no additive effect of Age Factor in the Log_Tax and Log_Price relationship.
RSS(Covariate)-RSS(Covariate+Factor)=0.1040 with 1 d.f.
The best estimate of the model variance comes from the complete model and is 1.5485/62 which is 0.02497 with 62 d.f. Therefore, the quotient 0.1040/0.02497 is equal to 4.1648 and therefore the Fisher statistic of (1 and 62) d.f. will be significant, specifically the p_value is 0.0457 (remember p_value=P(F(1,62) > 4.16)=0.0457) less than the usual reference 5% threshold (do your best just with statistical tables, an exact pvalue can not be obtained). Then there is evidence to reject the null hypothesis and therefore cars pay more tax when being new.

3. What is the multiple correlation between the response and its two predictors.

I am assuming to discuss the additive Ancova model explains (5.8851+0.1040) units of 7.5583 units, that is 0.7924 to 1 or 79.24% of target variability (coefficient of determination) and therefore the multiple correlation coefficient is the square root of the coefficient of determination and thus 0.8902.
If the interactive model is chosen, model explains 5.8851 units of 7.5583 units of target's variability, so 77.86% for coefficient of determination and its squared root for multiple correlation coefficient.

4. Is there statistical evidence to suggest that the relationship between the response and the covariate differs in new or old vehicles?

A contrast of the interaction between Covariate and Factor is requested: no alternative. We do this by applying Fisher test. RSS(Covariate+Factor)-RSS(Covariate*Factor)=0.0207 with 1 d.f. The best estimate of the model variance comes from the complete model and is 1.5485/62 = 0.02497 with 62 d.f. (if additive model variance estimate is used then it has been considered as almost correct). Therefore, the quotient 0.0207 /0.02497= 0.8288 and therefore the Fisher statistic of (1,62) d.f.

will be non-significant, specifically the p_value is 0.3661 (remember p_value=P(F(1,62) > 0.8288)= 0.3661, you are not supposed to give the exact figures, just whether H0 is rejected, or not) much higher than the usual reference 5% threshold. Then there is no evidence to reject the null hypothesis and therefore, **thus the linear relationship between price and taxes does not depend on vehicle age factor**.

5. There is one observation with a leverage greater than 0.1. Mark it on the bivariate diagram. What is the maximum acceptable value for leverage according to traditional results?

The maximum acceptable value is 2p/n, in the additive model it will be 6/66=0.091. If you have assumed an alternative model to discuss the threshold, then it is also correct. Obviously it must be a very large or very small observation in the explanatory variable log_Price. It looks like the observation of the lowest price, but graphically you don't have more elements: yes, you should say a priori that they can be the most expensive or the cheapest of each type of car, old and new by interpreting leverage as diagonal of the projection matrix. In fact it turns out that it is the observation of old car and more expensive price!

6. There is one observation with a studentized residual lower than -2.7 in the additive model. Mark it on the bivariate diagram.

The negative sign indicates that it is a vehicle that pays relatively few rates, so it points towards a new vehicle that is below 'its axis'. It is the observation corresponding to log_Price 6.8 and minimum log_rate within the group of new cars. Other reasonable possibilities are given as valid and are circled in the initial graph.

7. Are there any observations suspected of being influential data? What is the maximum reference value for the statistic you would use to formally answer the above question? Which is this statistic?

I have 2 candidates among the old vehicles, the cheapest (less than 6.4 in logPrice) and the second most expensive (approx. logPrice 7.5). There are authors who consider as a posteriori influential observations those that have an unusually high Cook's distance with respect to the magnitudes of the rest of the observations. Chatterjee and Hadi propose using a threshold of 4/(n-p), which in our case is 0.0635. Therefore, I would use a boxplot and decide according to a standard descriptive statistics criterion.

Name:
DNI/Passport:

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

Part 2-Problem 3 (6 points): All questions account for the same weight

Catalan Parliament elections in 2015, 2017 and 2021 results are considered in this exercise. Data has been obtained on https://eleccions.gencat.cat/ca/resultats-electorals/#/. Selected parties are those obtaining seats in the Catalan Parliament. Names for parties have been adapted to allow a comparison across years.

wing	independency	party	party.name	votes.15	seats.15	votes.17	seats.17	votes.21	seats.21
Left	no	**CatComú-Podem**	Comuns-Podem	367613	11	326360	8	195345	8
Right	no	**Cs**	Ciutadans-Partido de la Ciudadanía	736364	25	1109732	36	158606	6
Left	yes	**CUP**	Candidatura d'Unitat Popular	337794	10	195246	4	189924	9
Center	yes	**ERC**	Esquerra Republicana/ Junts pel Sí	0	0	935861	32	605581	33
Center	yes	**JUNTSxCAT**	Junts pel Sí/Cat	1628714	62	948233	34	570539	32
Right	no	**PP**	Partit Popular	349193	11	185670	4	109453	3
Left	no	**PSC**	Partit dels Socialistes de Catalunya (PSOE)	523283	16	606659	17	654766	33
Right	no	**VOX**	Vox	0	0	0	0	218121	11
Others	unknown	**Others**	Others	187235	0	85130	0	182510	0
Abstention	unknown	**Abstention**	Abstention	1380657	0	1161564	0	2739222	0

Let us focus on attitudes towards Catalonia becoming an independent state in the EU. New tables grouping by independency opinion according to party wing and Abstention have been elaborated and are shown below.

Catalan Parliament Elections		year		
wing	**independency**	**2015**	**2017**	**2021**
Center	**yes**	1,628,714	1,884,094	1,176,120
Left	**no**	890,896	933,019	850,111
Left	**yes**	337,794	195,246	189,924
Right	**no**	1,085,557	1,295,402	486,180
Others	**unknown**	187,235	85,130	182,510
Abstention	**unknown**	1,380,657	1,161,564	2,739,222
Total		**5,510,853**	**5,554,455**	**5,624,067**

Year	Abtention-Yes	Abstention-No	Census
2015	1380657	4130196	**5510853**
2017	1161564	4392891	**5554455**
2021	2739222	2884845	**5624067**
Total	**5281443**	**11407932**	**16689375**

Let us address firstly Abstention across years.

1. Determine binary logit null model parameter estimate for abstention proportion (m1) based on grouped data (wing factor is not considered, as yes/no are also grouped). Null deviance is 1140214 units.

Firstly, data have to be grouped in the proper way, to simplify calculus, and the table on the right has the convenient format. Average probability of abtention is 0.3165 (=5281443/16689375), odds are 0.4640 and logodds -0.7701.

$$(M1)\ log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta \rightarrow \hat{\eta} = -0.7701$$

Year	Abtention-**Yes**	Abtention-**No**	**Census**
2015	1380657	4130196	**5510853**
2017	1161564	4392891	**5554455**
2021	2739222	2884845	**5624067**
Total	**5281443**	**11407932**	**16689375**

2. Determine binary logit model statement and estimate parameters for the abstention proportion across years (m2).

$$(M2)\ log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta + \alpha_i \quad i = 1..3 \quad i = 1 \equiv 2015\ and\ \alpha_1 = 0$$
$$i = 1 \equiv 2015 \quad \hat{\eta} = log\frac{1380657}{4130196} = -1.0958$$
$$i = 2 \equiv 2017 \quad \hat{\alpha}_2 = \hat{\eta} + \hat{\alpha}_2 - \hat{\eta} = log\frac{1161564}{4392891} - log\frac{1380657}{4130196} = -1.3302 + 1.0958 = -0.2345$$
$$i = 3 \equiv 2021 \quad \hat{\alpha}_3 = \hat{\eta} + \hat{\alpha}_3 - \hat{\eta} = log\frac{2739222}{2884845} - log\frac{1380657}{4130196} = -0.0518 + 1.0958 = 1.0440$$

3. Address a deviance test to determine whether year factor is significant or not in the abstention proportion.

A deviance test stating HO:'Models (M1) and (M2) are equivalent' based of the asymptotic distribution of $\Delta Dev(M1, M2)$ as $\chi^2_{V=2}$ has to be addressed.

$$D = 2\sum_{i=1,..,3}\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right\} = 1140214 \approx \chi^2_{n-p=3-1=2}$$

Residual deviance for the null model (M1) is 1140214 and residual deviance for (M2) is 0 since it is a saturated model:

$$\Delta Dev(M1, M2) = Dev(M1) - Dev(M2) = 1140214 - 0 = 1140214$$

Thus $P(\chi^2_2 > 1140214) = 0$ and ***HO can be rejected, abstention proportion is not the same across years Available data.***

Independency position is taken as the response variable (categories are unknown, no and yes in that particular order) in the following questions. The results of fitting the additive multinomial logit model using wing and year factors are presented below. It has a pseudo-coefficient of determination (McFadden) of 0.905. By adding the interaction between wing and year factors, the logarithm of the likelihood goes up 0.0007416277 units:

NO vs Unknown	Estimates	YES vs Unknown	Estimates
Wing-Center	-10.975	**Wing-Center**	21.692
Wing-Left	-32.168	**Wing-Left**	-0.222
Wing-Right	0.262	**Wing-Right**	-0.445
Wing-Others	-31.441	**Wing-Others**	-18.942
Year-2017	2.559	**Year-2017**	-0.596
Year-2021	2.332	**Year-2021**	-0.529
constant	13.722	**constant**	-0.747
LogLik	**-1736.702**	**LogLik Null Model**	**-18332.160**
Explained Deviance	**33190.920**	**Residual Deviance**	**3473.404**

Name:

DNI/Passport:

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

4. Formally state the model. Detail the number of parameters of the additive and interactive models. What is the residual deviance of the wing + year additive model?

$$\log\left(\frac{\pi_{ijl}}{\pi_{ij1}}\right)=\eta_l+\alpha_{il}+\beta_{jl}\quad with \quad \begin{array}{l} i=1,\ldots 5\ \ wing\ \ baseline\ 1\equiv abstention \\ j=1,2,3\ year\ \ baseline\ 1\equiv 2015 \\ l=1,2,3\ \mathrm{Re}\,sponse\ \ baseline 1\equiv Unknow \end{array}$$

There are 2 logodd equations 1) No vs Unknown 2) Yes vs Unknown. For each equation the number of parameters are 1+4+2, thus 7x2=14 parameters. In case of wing*year interactive model, the number of parameters would be: 2*(1+4+2+8)=30 parameters. Residual deviance is twice minus the logLik function value Dev(M4)=2(-logLik(M4))=2 *(-1736.702)= 3473.404.

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

- In the case of logodd equation for No vs Unknown is a linear term that adds 2.56 for 2017 and 2.3 for 2021 to **2015-reference level, all else being equal.**
- In the case of logodd equation for Yes vs Unknown is a linear term that adds -0.5961068 for 2017 and -0.528918 for 2021 to 2015-reference level, all else being equal. So, logodds are decreased **compared to the base year (2015) all else being equal.**
- In the case of odd for No vs Unknown the effect is multiplicative by 13=exp(2.56) for 2017 and 10.3=exp(2.33) for 2021 **with respect to 2015**-reference level odds, **all else being equal.** So, relative probability of No vs Unknown increase by 1192% and 939% for 2017 and 2021 **with respect to 2015** base-year **all else being equal.**
- In the case of odd for Yes vs Unknown the effect is multiplicative by 0.55=exp(-0.6) for 2017 and 0.59=exp(-0.53) for 2021 with respect to 2015-reference level odds, **all else being equal.** So, relative probability of No vs Unknown decrease by 45% and 41% for 2017 and 2021 **with respect to 2015** base-year **all else being equal.**

6. What are the predicted probabilities for the response categories for individuals voting parties on the Left wing in 2017?

i=3 (left), j=2 (2017) and l=1 (Unknown), 2 (No) and 3 (Yes) refer to index meaning in model statement:

$$log\left(\frac{\pi_{322}}{\pi_{321}}\right) = 13.72197 - 32.1675706 + 2.5590272 = -15.89 \rightarrow \frac{\pi_{322}}{\pi_{321}} = 1.26e-07$$

$$log\left(\frac{\pi_{323}}{\pi_{321}}\right) = -0.74707 - 0.2222275 - 0.5961068 = -1.57 \rightarrow \frac{\pi_{323}}{\pi_{321}} = 0.209$$

$$\pi_{321} = \frac{1}{1+\frac{\pi_{322}}{\pi_{321}}+\frac{\pi_{323}}{\pi_{321}}} = 0.827$$

$$\pi_{322} = \pi_{321}\frac{\pi_{322}}{\pi_{321}} = 1.042606e-07$$

$$\pi_{323} = \pi_{321}\frac{\pi_{323}}{\pi_{321}} = 0.173$$

7. Is there any evidence to affirm that party wing effect is different according to year?

This question involves adding interactions between wing and year. LogLik increases by 0.0007416277 units, thus $\Delta Dev = Dev(wing + year) - Dev(wing * year) = 2\cdot 0.0007416277 = 0.00182$, to be compared to $\chi^2_{16} \rightarrow P(\chi^2_{16} > 0.00182) \cong 1$. Tables are not needed, H0 'Both models are equivalent' can not be rejected, thus interactions are not required.

Coefficients	Estimates (latent)
Wing-Center	-0.163
Wing-Left	-23.478
Wing-Right	-2.570
Wing-Others	-43.722
Year-2017	-0.767

Year-2021	-0.951
Constant No|Unknown	-21.734
Constant Unknown|Yes	-7.937
LogLik	**-52511.41**
LogLik Null Model	**-18342.03**
Residual Deviance	**105022.82**
Null Deviance	**36684.05**

8. Formulate the model. Detail the number of parameters of the additive and interactive models. Use level order as no, unknown and yes in all the sections.

$$\log\left(\frac{\gamma_{ijl}}{1-\gamma_{ijl}}\right) = \eta_1 - \alpha_i - \beta_j \quad with \quad \begin{array}{l} i = 1,\ldots 5 \;\; wing \;\; baseline \; 1 \equiv abstention \\ j = 1,2,3 \;\; year \;\; baseline \; 1 \equiv 2015 \\ l = 1,2 \;\; \mathrm{Re}sponse \;\; baseline 1 \equiv Unknow \end{array}$$

Additive model number of parameters is p=2+4+2=8 and interactive model one p'=8+8=16.

9. Interpret the effect of year in terms of proportional odds and latent variable.

In terms of latent variable analysis, coefficients are directly the ones provided and propension to independency decreases in the logodds scale for all years: -0.7669607 for 2017 and -0.9512354 for 2021. 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, 2017 year has an effect on the propension scale of moving no No or Unknown 0.423 standard deviations to the left in the propension scale with respect to the base year and 2021 pushes to the left 0.524 standard deviations with respect of 2015 reference year.

Under the proportional odds point of view, 2017 year increases by 115% odds of No vs (Unknown or Yes) and odds of No or Unknown versus Yes, all else being equal. And 2021 year increases by 159% the relative probabilities of No vs (Unknown or Yes) and odds of No or Unknown versus Yes, all else being equal.

$$log\left(\frac{\gamma_{i21}}{1-\gamma_{i21}}\right) = log\left(\frac{\pi_{i21}}{\pi_{i22}+\pi_{i23}}\right) = -21.73 + 0.767 + \alpha_i$$

$$log\left(\frac{\gamma_{i21}}{1-\gamma_{i21}}\right) \Big/ log\left(\frac{\gamma_{i11}}{1-\gamma_{i11}}\right) = log\left(\frac{\gamma_{i22}}{1-\gamma_{i22}}\right) \Big/ log\left(\frac{\gamma_{i21}}{1-\gamma_{i21}}\right) = \exp(0.767) = 2.153 \rightarrow 100(2.153-1) = 115.3\% \uparrow$$

$$log\left(\frac{\gamma_{i31}}{1-\gamma_{i31}}\right) \Big/ log\left(\frac{\gamma_{i11}}{1-\gamma_{i11}}\right) = \exp(0.951) = 2.59 \rightarrow 100(2.59-1) = 159\% \uparrow$$

10. What are the predicted probabilities for the response categories for individuals voting parties on the left wing in 2017 according to the ordinal proposal?

i=3 (left), j=2 (2017), refer to index meaning in model statement:

$$log\left(\frac{\gamma_{321}}{1-\gamma_{321}}\right) = log\left(\frac{\pi_{321}}{\pi_{322}+\pi_{323}}\right) = -21.73 + 23.48 + 0.767 = 2.51 \rightarrow \gamma_{321} = \frac{exp(2.51)}{1+exp(2.51)} = 0.925$$

$$log\left(\frac{\gamma_{322}}{1-\gamma_{322}}\right) = log\left(\frac{\pi_{321}+\pi_{322}}{\pi_{323}}\right) = -7.94 + 23.48 + 0.767 = 16.31 \rightarrow \gamma_{322} = \frac{exp(16.31)}{1+exp(16.31)} = 0.9999 \rightarrow \pi_{322} = \gamma_{322} - \gamma_{321} = 0.0752$$

$$\pi_{323} = 1 - \gamma_{322} = 0.000001$$

So, No probabilily 0.925, Unknown probability 0.0752 and Yes probability 0.000001

11. Compare the nominal/ordinal additive propositions according to Akaike's criterion. Have you seen any abnormal outcome?

Loglikelihood function for the null model is greather than the one for the additive mode (wing+year) in the ordinal proposal: this does not make sense and the optimization process does not converge. Something wrong is happening. Mininum AIC is obtained by the nominal proposal.

AIC(nominal)=2*(-logLik(nominal) + p(nominal))= 2(1736.702+14)= 3501.404

AIC(ordinal)= 2*(-logLik(ordinal) + p(ordinal))= 2(52511.41+8)= 105038.82