**MASTER IN DATA SCIENCE (FIB-UPC).**

**ACADEMIC YEAR 22-23 Q1** – **FINAL EXAM**

**Statistical Inference and Modelling (SIM)**

**Date: 11/Jan/2023 15-18:30h 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 inn Final Exam Allowed Documents folder on the ATENEA 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.

**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.**

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

**Suppose that we have reason to believe that the readings x1, x2, . . . , x16 obtained from an experiment were a random sample from a N(μ, σ=2) distribution.**

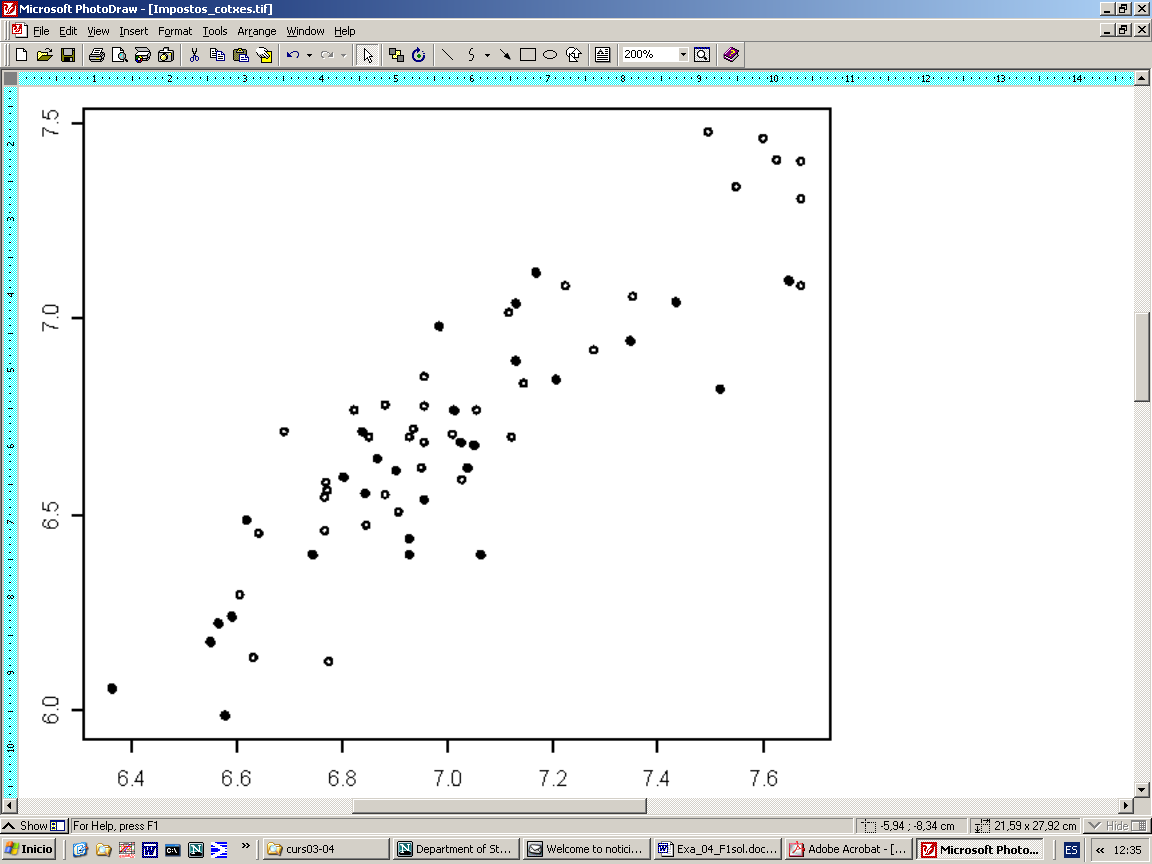
1. We wish to test H0 : μ = μ0 = 36.0 versus H1 : μ < 36.0. If the observed value of the sample mean is 34.4, what would be the outcome of the test at the 5% significance level?
2. We wish to test H0 : μ = μ0 = 36.0 versus H1 : μ ≠ 36.0. If the observed value of the sample mean is 34.4, what would be the outcome of the test at the 1% significance level?

**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.**

1. Test at the 1% and 5% level whether μ = 40 or not.
2. Determine a 95% two-sided interval for population variance.

**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).



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.
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.
3. What is the multiple correlation between the response and its two predictors?
4. Is there statistical evidence to suggest that the relationship between the response and the covariate differs in new or old vehicles?
5. There is one observation with leverage greater than 0.1. Mark it on the bivariate diagram. What is the maximum acceptable value for leverage according to traditional results?
6. There is one observation with a studentized residual lower than -2.7 in the additive model. Mark it on the bivariate diagram.
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?

**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 Socialistes de Catalunya (PSC-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.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

**Let us address 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.**
2. Determine binary logit model parameter estimates for abstention proportion across years gross effect (m2).
3. Address a deviance test to determine whether year gross factor is significant or not in the abstention proportion.

**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 |
| Overall GoF results: |  |  |  |
| LogLik | **-1736.702** | **LogLik Null Model** | -18332.160 |
| Explained Deviance | **33190.920** | **Residual Deviance** | - |

1. Formally state the model. Detail the number of parameters of the additive and interactive models. What is the residual deviance of the nominal wing + year additive model?
2. Interpret the effect of year on the outcome in terms of logodds and relative probabilities (odds) in the nominal wing+year additive model.
3. What are the predicted probabilities for the response categories for individuals voting parties on the left wing in 2017 according to the nominal wing + year additive model?
4. Is there any evidence to affirm that party wing effect is different according to year?

|  |  |
| --- | --- |
| Coefficients | Estimates (latent point of view) |
| 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 |

1. Formally state 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.
2. Interpret the effect of year in terms of proportional odds and latent variable according to wing +year ordinal additive model.
3. What are the predicted probabilities for the response categories for individuals voting parties on the left wing in 2017 according to the ordinal wing+year additive proposal?
4. Compare the nominal/ordinal additive proposals according to Akaike's criterion. Can you see any abnormal outcome?