Instructions:

- Answer the following questions using SAS and provide the code you used to perform the analysis in a separate file (`.txt` extension, `utf8` encoding).
- Your report must be submitted as a PDF file and should not exceed 15 pages; any additional page will be ignored. Be brief, but precise; only include relevant output.
- When you perform an hypothesis test, always write down the null and alternative in terms of the model parameters, define the latter, choose an appropriate test statistic, report its numerical value, its null distribution, the conclusion of the test and interpret your results in the context of the problem.
- Errors are penalized even if they are not directly related to the question.
- (Sub)-questions marked with a star ($\star$) should be completed in groups of two or three. Provide the graphs and the results only in one of the two reports and include the name of your teammates in each report.

1.1 **Graphics** ($\star$). Choose two graphs online an publish a link to the source on Piazza. You cannot use one of the graphs analyzed in class, nor one chosen by your teammates (first come, first served).[1] Marks will be given for original choices.

For each graph, briefly discuss the following elements:

- Summarize in your own words the narrative of the graph.
- What type of graph has been used (geometry): is this choice appropriate in the context?
- Identify the different type of variables and mapping (x and y-axis, shape, color, etc.)
- Identify strengths and weaknesses of the visual.

1.2 **Calculating the power of a statistical test**

The SAS program `power.sas` contains code by Rick Wicklin from SAS Institute Inc. to perform a simulation study for computing the power of the two-sample t-test, which is the same as that of the binary variable indicating group in a simple linear regression model.

(a) Briefly explain in your own words the steps of the simulation study.

(b) Plot and comment on the graph produced with parameters $n_1 = n_2 = 10$, $\sigma = 1$ and $B = 10000$.

(c) Vary the number of simulations from $B = 100$ to $B = 10000$. What do you notice when the number of simulations is small? Explain why this effect vanishes as the number of simulations increase.

(d) Modify the code so that the size of the groups is (a) $n_1 = 10, n_2 = 30$ and (b) $n_1 = 20, n_2 = 20$. In which of these two scenarios is the power highest and why?

(e) Modify the code to simulate $n = 20$ observations in each group from a normal distribution with different standard deviations, $\sigma_1 = 1, \sigma_2 = 5$. Report the estimated level of the test along with the number of simulations and a 90% confidence interval. Explain how you derived the latter.

1.3 **Linear model for `diamonds`** The `diamonds` contains the price of 1000 cut diamonds along with the following variables:

- `price`: price (in USD)
- `carat`: weight (in carats)
- `cut`: quality of the cut (`fair`, `good`, `very good`, `premium`, `ideal`)
- `color`: color, either `colorless` (DEF) or `near colorless` (GHIJ)
- `clarity`: clarity , ranging from `I1` (worst), `SI2`, `SI1`, `VS2`, `VS1`, `VVS2`, `VVS1` to `IF` (best)
- `length`: length (in mm)
- `width`: width (in mm)
- `depth`: depth (in mm)
- `totdepth`: total depth `2depth/(width+length)` (in percentage)
- `table`: width of top of diamond relative to widest point

[1] Some suggestions: Twitter feeds of the BBC, the Washington Post, the New York Times, etc., magazine or official agencies like Statistic Canada or the US Census Bureau. Avoid generic time series (e.g., stock market price).

Compiled 14/10/2020 at 10:32

(a) (⋆) Perform an exploratory data analysis (one page max of text). Summarize the key elements necessary for adequately interpreting the data. Address the following points:
- Are there outliers or data entry errors in the database?
- What is the relation between price and carat?
- Which, if any, of the explanatory variables are correlated between themselves?
- We could add dimensions (`length`, `width`, `depth`, `totdepth`, `table`) as covariates in a linear model: would this be logical or not?

(b) Fit a linear model for the price of diamonds with `carat`, `cut`, `color` and `clarity` as explanatory variables. Use `fair`, `near colorless`, `I1`, as baseline categories.
i. Interpret the parameters associated to `cut` (`very good`) and `carat`.
ii. Interpret the coefficient for `color` (`colorless`). Is the value of the estimate logical? Explain.
iii. Is the increase in price the same for each additional level of clarity?
iv. Produce graphical diagnostics of residuals and comments on the validiy of the linear model assumptions.

(c) Suppose we model instead ln(`price`), with the same set of explanatory variables.
i. Should one also log-transform the explanatories? Justify your answer and fit the model with the (un)transformed explanatories, depending on your answer.
ii. Interpret the parameters for `cut` (`very good`) and `carat` (on the original scale, meaning changes in USD).
iii. Which of the additive model or the multiplicative model (log-transformed) seems most adequate? Justify your answer using graphics, tests or other criteria.

1.4 The `renfe` data contains information about 10 000 train ticket sales from Renfe, the Spanish national train company. The data include:
- `price`: price of the ticket (in euros);
- `dest`: binary variable indicating the journey, either Barcelona to Madrid (`0`) or Madrid to Barcelona (`1`);
- `fare`: categorical variable indicating the ticket fare, one of `AdultoIda`, `Promo` or `Flexible`;
- `class`: ticket class, either `Preferente`, `Turista`, `TuristaPlus` or `TuristaSolo`;
- `type`: categorical variable indicating the type of train, either Alta Velocidad Española (`AVE`), Alta Velocidad Española jointly with TGV (parternship between SNCF and Renfe for trains to/from Toulouse) `AVE-TGV` or regional train `REXPRESS`; only trains labelled `AVE` or `AVE-TGV` are high-speed trains.
- `wday` integer denoting the week day, ranging from Sunday (`1`) to Saturday (`7`).

Fit a linear regression to explain the `price` of high-speed trains as a function of ticket `class`, `fare`, an interaction between `class` and `fare` and a dummy variable indicating whether the train travels during a week-end or not. Use `Flexible`, `Preferente`, and weekday as baselines.

(a) Write down the equation of the postulated theoretical model.

(b) Is the interaction term between `class` and `fare` statistically significant?

(c) Report the estimated coefficient and interpret the parameters associated to `class` and `fare` (including the interaction terms between the two).

(d) Explain why we would not consider the F test for `class` in the model with interaction: what does this test represent in the context?

(e) Test for the global significance of the model.

(f) Predict the price of an `AVE` at `Promo` fare, class `Turista` for a Saturday and give a 90% prediction interval for the latter.

(g) Produce diagnostic plots of residuals and comment on the validity of the linear model assumptions.

1.5 **To be handed in along with Assignment 2** We consider a simple Poisson model for the number of daily sales in a store, which are assumed independent from one another. Your manager tells you the latter depends on whether

Compiled 14/10/2020 at 10:32

there are sales or not. The mass function of the Poisson distribution is

$$\mathsf{P}\left(Y_i = y_i \mid x_i\right) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \qquad y_i = 0, 1, \ldots$$

where we model $\lambda_i = \exp(\beta_0 + \beta_1 \texttt{sales}_i)$, where $\texttt{sales}_i$ is a binary indicator equal to unity during sales and zero otherwise.

(a) Derive the maximum likelihood estimator of (β_0, β_1). *Hint: maximum likelihood estimators are invariant to reparametrization. It is easier to obtain them for both subsamples of sale period/regular and then map the estimates to* (β_0, β_1)

(b) Calculate the maximum likelihood estimates for a sample of size 12, where the number of transactions outside sales is $\{2;5;9;3;6;7;11\}$, and during sales, $\{12;9;10;9;7\}$.

(c) Calculate the observed information matrix and use the latter to derive standard errors for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and a 95% confidence interval for the parameters.

(d) Your manager wants to know if the daily profits during sales are different from those outside of the sale periods. She calculates that the average profit during sales is \$20 per transaction, compared to \$25 normally. Test this hypothesis using a likelihood ratio test. *Hint: write the null hypothesis of equal profit in terms of the model parameter.*

Compiled 14/10/2020 at 10:32