PIE2. First Deliverable. 2018-19

Note:

In this first deliverable you have to perform the analysis that is explained in the STATEMENT section.
You do not have to deliver any report or file, you just have to answer the questions that appear in the QUESTIONNAIRE section by means of ATENEA.
Important to know that you have as many opportunities as you want, in the sense that if you change your opinion and you want to change the answer, you can do it, until the deliverable deadline

STATEMENT

We have the hight, $\boldsymbol{H}$, of ficus plants and the number of days it has been planted, ***Days***. The data ($Days$ and H) appear in the file that has your name in ATENEA, in the folder corresponding to "Dades 1r Entregable".

This data set constitutes an example of data such that requires to transform the response variable by means of the logarithm, and to consider the regression $log\,(H) \sim \alpha + \beta \cdot Days$. But as you will see, the model doesn't verify the homoscedasticity hypothesis. Moreover, as you will see, the non-linear regression model $H \sim e^{\alpha+\beta \cdot Days}$ doesn't verify the homoscedasticity property neither (at least it is not clear).Thus the analysis of this model point out the need to use generalized linear models and this is what will be done in the second deliverable.

In all the exercise the significance level is set to be equal to $\alpha = 0.05$.

The models that we are going to work with are defined in what follows:

- ***M1***: The regression line of H with respect to the variable $Days$.
- ***M2***: The regression parabola of H with respect to $Days$ (i.e, in the linear part it appear the explanatory variables $Days$ and $Days^2$).
- ***M3***: The regression line of $Log(H)$ with respect to $Days$. This transformation is useful when one wants to stabilize the variances, and the variances are approximately a quadratic function of the mean $Var(H_i) \simeq (\mu_i)^2$.
- ***M4***: The regression line of $\sqrt{H}$ with respect to $Days$. This transformation is useful when one wants to stabilize the variances and the variances are approximately a linear function of the mean $Var(H_i) \simeq \mu_i$.
- ***M5***: The non-linear regression model defined as:
  $H_i = e^{\alpha+\beta\, Days_i} + e_i$, were e_i are iid r.v's with distribution $N\,(0, \sigma^2)$.

To perform this non-linear regression the procedure ***nls*** of R is required. Please for information about this procedure use the R help. You also have some examples of application in the scripts associated to exercises: T1-E05 and T2-E01.

For each one of the models presented, we are going to focus in the goodness-of-fit of the model and in checking if the linear model hypothesis are verified. In particular special attention will be made in the homoscedasticity hypothesis. Moreover, we are going to be interested in predicting the hight of the ficus when planted ($Days = 0$) and after 105 and 150 days of being planted.

For each one of the models:

1. Fit the data, compute the parameter estimations and interpret them.

2. Perform the residual analysis. In particular plot the residuals vs fitted. In the case of the first four models (the linear ones), also perform the residual analysis using the standardized residuals.

3. Let us define $FDays$ as the variable $Days$ considered as a Factor. This can be done, because there are sufficient number of observations for each value of the $Days$ variable. For the first four models, compare the fits obtained with the ones obtained if at each model it is added the variable $FDays$ as an extra explanatory variable. Do it by means of the anova test $anova\,(model1, model2)$. In particular, for the first model considered, we are asking you to compare the model $H \sim Days$ with the model $H \sim Days + FDays$.

4. As we said before, given that we have a large number of observations for each value of the variable $Days$, perform the Levene's test test to compare the variances in the different groups and see if the homoscedasticity properly may be assumed or not. Information about the Levene's test may be found in the R help.

5. Estimate the mean and the standard deviation associated to the variable H when $Days$ is equal to $0, 105$ and 150. Remind that the variance of the response variable is constant and equal to the error variance. That is, we ask you to estimate:

$$E(H|Days = a) \text{ and } \sqrt{Var\,(H|Days = a)},$$

for $a = 0$, 105 and 150.

Hint: I*n the models where the response variable has been transformed by means a function* $g\,(H)$*, one has that if:*

$$m = \widehat{E}(g\,(H)\,|Days = a) \text{ and } s = \sqrt{\widehat{Var}\,(g\,(H)\,|Days = a)},$$

then, denoting $f\,(x) = g^{-1}\,(x)$*, one has that*

$$E(H|Days = a) = f\,(m) \text{ and } \sqrt{\widehat{Var}\,((H)\,|Days = a)} = s \cdot |f'\,(m)|\,.$$

Compare the values that you have obtained with the ones that appear in the file ES.csv.

6. Compare the results obtained with the five models.

QUESTIONNAIRE

1. *Model: M1 Ap: 1.* The intercept's estimate is (3 dec.):
2. *Model: M1 Ap: 1.* The estimation of the residual standard error is (3 dec.):
3. *Model: M1 Ap: 2.* The number of |studentized residuals| greater than 3 is:
4. *Model: M1 Ap: 4.* Does it exist homoscedasticity?
   (a) Yes, it exists
   (b) We are not very sure about it, but we do not reject that it exists.
   (c) No, sure it doesn't exist.
   (d) We are not very sure about it, but we do not reject that it doesn't exist.
5. *Model: M1 Ap: 3.* The value of the contrast statistics of the anova test is equal to (2 dec.):
6. *Model: M1 Ap: 5.* The estimation of $E\left[H_{|Days=105}\right]$ is (3 dec.):
7. *Model: M2 Ap: 1.* Is the coefficient of $Days$ equal to zero?
   (a) Yes, sure it is equal to zero.
   (b) We are not very sure about it, but we accept that it is equal to zero.
   (c) No, sure it isn't equal to zero.
   (d) We are not very sure about it, but we accept that it isn't equal to zero.
8. *Model: M2 Ap: 1.* The estimation of the residual standard error is (3 dec.):
9. *Model: M2 Ap: 2.* Which is the first Day such that the |studentized residual| is larger than 2? answer zero in case where it never has a value larger than 2.
10. *Model: M2 Ap: 4.* The value obtained of the contrast statistic associated to the Levene's test is equal to (3 dec.):
11. *Model: M2 Ap: 3.* The *p-value* of the anova test is equal to (if p-value>0.001 4 dec. else in scientific notation):
12. *Model: M2 Ap: 5.* The estimation of $S\left[H_{|Days=105}\right]$ is (3 dec.):
13. *Model: M3 Ap: 1.* The slope's estimate is (3 dec.):
14. *Model: M3 Ap: 1.* The estimation of the residual standard error is (3 dec.):
15. *Model: M3 Ap: 2.* Which is the last day in which the |studentized residual| is larger than 2? Answer 0 if it never gets larger than 2.
16. *Model: M3 Ap: 4.* Does it exist homoscedasticity?
    (a) Yes, it exists
    (b) We are not very sure about it, but we do not reject that it exists.

(c) No, sure it doesn't exist.

(d) We are not very sure about it, but we do not reject that it doesn't exist.

17. *Model: M3 Ap: 3.* Does the model appropriately fit the mean of H?

    (a) "Yes", sure.

    (b) We are not very sure about it, but we accept "yes".

    (c) "No", sure.

    (d) We are not very sure about it, but we accept "no".

18. *Model: M3 Ap: 5.* The estimation of $S\left[H_{|Days=150}\right]$ is (3 dec.):

19. *Model: M4 Ap: 1.* The standard error of the intercept is (4 dec.):

20. *Model: M4 Ap: 1.* The estimation of the residual standard error is (3 dec.):

21. *Model: M4 Ap: 2.* Which is the last day in which the |studentized residual| is larger than 2? Answer 0 if it never gets larger than 2.

22. *Model: M4 Ap: 4.* Does it exist homoscedasticity?

    (a) Yes, it exists

    (b) We are not very sure about it, but we do not reject that it exists.

    (c) No, sure it doesn't exist.

    (d) We are not very sure about it, but we do not reject that it doesn't exist.

23. *Model: M4 Ap: 3.* Does the model appropriately fit the mean of H?

    (a) "Yes", sure.

    (b) We are not very sure about it, but we accept "yes".

    (c) "No", sure.

    (d) We are not very sure about it, but we accept "no".

24. *Model: M4 Ap: 5.* The estimation of $S\left[H_{|Days=0}\right]$ is (3 dec.):

25. *Model: M4 Ap: 5.* The estimation of $S\left[H_{|Days=150}\right]$ is (3 dec.):

26. *Model: M5 Ap: 1.* The standard error of the intercept is (4 dec.):

27. *Model: M5 Ap: 1.* The estimation of the residual standard error is (3 dec.):

28. *Model: M5 Ap: 4.* Does it exist homoscedasticity?

    (a) Yes, it exists

    (b) We are not very sure about it, but we do not reject that it exists.

    (c) No, sure it doesn't exist.

    (d) We are not very sure about it, but we do not reject that it doesn't exist.

29. *Model: M5 Ap: 5.* The estimation of $E\left[H_{|Days=0}\right]$ is (3 dec.):

30. Ap: 6. Based on the log-likelihood, which of the five models do you thing it is the best?

    M1 M2 M3 M4 M5

31. Ap: 6. Based on the tests of "Ap: 3.", which of the four linear models (M1-M4) do you think is the best for modeling the mean of the response variable?

    M1 M2 M3 M4

32. Ap: 6. Which of the five models do you think that better verifies the homoscedasticity hypothesis?

    M1 M2 M3 M4 M5

33. Ap: 6. Which of the five models fits better $E\left[H|_{Days=0}\right]$?

    M1 M2 M3 M4 M5

34. Ap: 6. Which of the five models fits better$S\left(H|_{Days=0}\right)$?

    M1 M2 M3 M4 M5

35. Ap: 6. For the five models, which is the smallest estimation of $S\left(H|_{Days=105}\right)$?