

02403: Assignment 3

two sample t-test, Anova

This is the third assignment in 02403, the main topics are two sample t-test and Anova.

Data:

A data set from measurements in Skive fjord is used to exemplify the methods you should implement in this exercise. The data set contain the following variables

year: The year of the measurement.

month: The month within the year.

chl_a: Average chlorophyll-a concentration in the month and year.

temp: Average water temperature in the month and year.

gr: Average global radiation (i.e. sunlight) in the month and year.

vmp: Water frame work directive, these are different legislations that are effective (given as integers from 0 to 3).

lchl_a: natural logarithm of chl_a.

The data set is also used in the lectures to exemplify different parts of the theory.

Hand in:

The answers should be kept short and concise, and may include Python code and output, but should be handed in as a single pdf-file. Some answers involve writing Python functions and in the case it should be explained what the function does.

Data for this assignment:

In this assignment we will consider lchla in the month of July under vmp=0 and vmp=2. You may extract a relevant data frame by

```
Skive = pd.read_csv("../data/skive.csv", sep=';')
dat = pd.DataFrame(Skive[Skive["month"]==7])
dat = pd.DataFrame(dat[(dat["vmp"]==0) | (dat["vmp"]==2)])
dat = pd.DataFrame(dat)
dat = dat.reset_index(drop=True)
dat
```

	year	month	chla	temp	gr	vmp	lchla
0	1982	7	0.02350	21.35	226.45	0	-3.750755
1	1983	7	0.03025	17.45	226.45	0	-3.498259
2	1984	7	0.01267	17.69	226.45	0	-4.368518
3	1985	7	0.02867	17.31	226.45	0	-3.551904
4	1986	7	0.02133	17.10	226.45	0	-3.847641
5	1987	7	0.00700	16.55	226.45	0	-4.961845
6	1999	7	0.01508	19.28	236.54	2	-4.194386
7	2000	7	0.00950	16.86	194.77	2	-4.656463
8	2001	7	0.00908	19.12	240.83	2	-4.701681
9	2002	7	0.00758	19.29	210.76	2	-4.882242
10	2003	7	0.01255	19.63	218.59	2	-4.378035

In this assignment, we consider the two sample t-test as a general linear model, i.e.

$$Y = X\beta + \epsilon; \quad \epsilon \sim N(\mathbf{0}, \sigma^2 I), \quad (1-1)$$

where $\beta = [\beta_1, \beta_2]^T$. We want to test whether the mean of lchla is the same under vmp=0 and vmp=2.

Questions:

- a) **Two sample t-test - Design matrix:** For the problem outlined above explicitly write down the design matrix for the three cases

Naive: $\hat{\beta}_i$ is the average in group i.

Stand: $\hat{\beta}_1$ is the average in group 1, and $\hat{\beta}_2$ is the difference between the average in the two groups.

Ortho: $\hat{\beta}_1$ is the average of all observations and the design is orthogonal.

Are "Naive" and "Stand" orthogonal designs? What is the interpretation of $\hat{\beta}_2$ in the design "Ortho"?

- b) **Two sample t-test - design matrix, implementation:**

1. Make a Python function that takes a pandas data frame, a column name as input and returns a design matrix where the parameter estimates are as the method "Naive".
2. Update the function such that you can choose between different interpretations of β , specifically include "Stand" and "Ortho".

The final function should be something like

```
X = two_samp_t_design(xnam, approach, data)
```

Where X is a design matrix, xnam is the column that defines the grouping, approach is the approach (naive, standard, and orthogonal).

- c) **Two-sample t-test – Estimation:** Update the one sample t-test estimation function from assignment 2 to include the two sample t-test, i.e. you should have a function with the call

```
fit = my_ols(y_nam, x_nam, approach, data)
```

where data is a Pandas dataframe, approach is one of the approaches defined above, x_nam is the grouping (here vmp) and y_nam is the response variable (here lchla). the function should return: the observations, the design matrix, the parameter estimates, the estimated standard deviation,

standard errors, observed teststatistics and p -values (for $H_0 : \beta_i = 0$), number of observations, degrees of freedom, date, x_nam , and y_nam . Make a function that summarise the results from `my_ols`, the call should be something like

```
print_my_ols(fit)
```

resulting in a print out

```
Summary table
Estimate  Std.Error    t_obs    p-val    CI_low  CI_high
beta_hat1   se1      t_obs1  p_value1 95% confidence interval
beta_hat2   se2      t_obs2  p_value2 95% confidence interval
Residual standard deviation
sigma_hat
```

you should test it with at least one of the design matrices define above.

- d) **Two-sample t-test – Verification:** Compare your results from the previous question with `stats.ttest_ind_ind(y1,y2,equal_var=True)`. I.e. with

```
fit1 = smf.ols('lchla ~ vmp', data=dat).fit()
print(fit1.summary(slim=True))
fit2 = smf.ols('lchla ~ -1 + vmp', data=dat).fit()
print(fit2.summary(slim=True))
```

what design matrix set up is used in each of the models? You can extract the design matrix by `fit1.model.exog`.

- e) **Two-sample t-test – Projections (Optional and possibly difficult):** Consider the Corollary 9.29 when the two-sample t-test is presented as projection in a general linear model. Answer the following questions. Use the geometric interpretation.

Hint: sketch the appropriate drawing from Example 9.32 for your understanding, it does not have to be included in the answer.

1. From what space to what subspaces are the projections H_0, H_1 ? what

are the dimensions of those subspaces? Do the dimensions of those subspaces depend on n_1, n_2 ?

2. Are $Y^T H_0 Y$ and $Y^T H_1 Y$ independent?

3. From what space to what subspace are the projections $H_1 - H_0$? and $I - H_1$?

4. What are the distributions of $H_0 Y$ and $H_1 Y$?

5. What improvement does $Y^T (H_1 - H_0) Y$ represent? Hint: compare the design matrices that the projection matrices are built from.

6. what assumptions are made in equations (9-142), (9-143), (9-144) such that the appropriate quantities follow chi squared distributions?

f) **Anova:** Based on the Corollary 9.29, update the anova function from assignment 2 to include the two sample t-test, with the hypothesis $H_0 : \beta_1 = \beta_2 = 0$, and $H_1 : \beta_1 = \beta_2$. Again test using different parametrizations of the design matrix, and comment on differences. Compare your results with the function `sm.stats.anova_lm(fit1, typ=1)`.

g) **Residual analysis:** Update the residual plot function from assignment 2 to include the fit from the two-sample t-test, the function produce boxplot of residual from each group and a common qq-plot.

Extra (might be difficult): You may consider

- Expanding with a leverage plot.
- to include the option of choosing between the raw, standardized and studentized residuals,
- The autocorrelation plot between the selected residuals and test of the lag-1 autocorrelation being zero.

Use the implemented function to perform residual diagnostics on the implemented linear model. Comment briefly on the results, are the assumptions of the model satisfied?