

02403: Assignment 2

One sample t-test, Anova tables, and residual analysis

This is the second assignment in 02403, the main topics are one sample t-test, Anova tables, and residual analysis.

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 use the monthly difference in temperature between 1998 and 2006. You may extract the data by

```
## Data
Skive = pd.read_csv("../data/skive.csv", sep=';')
y = Skive['temp']
year = Skive['year']
y1 = np.array(y[year==1998])
y2 = np.array(y[year==2006])
d = y1 - y2
```

The following questions on the one sample t-test and the data in d is used as the example data. The questions below (and in the next assignment) focus on the the general linear model description of data, i.e.

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

where in our case we have $d = \mathbf{y}$.

Questions:

- a) **One sample t-test:** For the concrete example given above, explicitly write down the design matrix $\mathbf{X}$, and find
- $\hat{\beta}$, $\hat{\sigma}^2$, and the standard error of $\hat{\beta}$.
 - The observed test-statistics and p -value for the hypothesis $H_0 : \beta = 0$.
 - The 95% confidence interval for β .

- b) **One sample t-test, implementation:** Write a function that takes a vector of observations as input. Running the function should result in a print out of the form

```
Summary table
      Estimate Std. Error    t_obs    p_val    CI_low    CI_high
0  beta_hat    se          t_obs    p_value 95% confidence interval
Residual standard deviation
sigma_hat
```

Further the function should return the design matrix, the observations, the parameter estimate and the residual standard deviation. I.e. you should be able to write something like

```
fit = my_one_sample_t_test(y)
```

- The test statistics (t_{obs}) and the p -value (p_{val}), should be for the null hypothesis $\mu = 0$. Further the function should return the design matrix, and the observation.
 - Test the function using the monthly difference in temperature between 1998 and 2006 and compare with the result from `stats.ttest_1samp`.
 - Extra: You may extend the function such that it can take a null-hypothesis different from zero as input, and also you may have an input that is the confidence level (α) (i.e. changing the confidence interval).
- c) **ANOVA table:** Referring to Corollary 9.27, write a Python function which takes the output from the function produced in the previous question and prints an analysis of variance table (ANOVA-table), i.e.

	SS	MS	F_{obs}	p -value
Effect	$\mathbf{Y}^T \mathbf{H} \mathbf{Y}$	$SS(\text{Effect})$	$\frac{MS(\text{Effect})}{MS(\text{Residual})}$	$P(F > F_{\text{obs}})$
Residual	$\mathbf{Y}^T (\mathbf{I} - \mathbf{H}) \mathbf{Y}$	$\frac{SS(\text{Residual})}{n-1}$		

I.e. you should be able to write

```
my_one_sample_t_aov(fit)
```

and get the anova table above.

Compare the result with the t-test conducted in the previous question (you may use `np.outer(X, X)` to calculate $\mathbf{X}^T \mathbf{X}$).

- d) **Residual analysis:** Make a function that takes the `fit` as input. The function should produce the following plots: A box plot of the residuals, and a qq-plot of the residuals. I.e. you should be able to write

```
my_dianostic(fit)
```

Use the function on the example from the previous questions.

- e) **Autocorrelation (Optional and possibly difficult):** Expand the dianostic function from question d) such that setting an option `autocorr = True`, result in at least one of the following additions

1. Plot of residuals at time t as a function of residuals at time $t - 1$.
2. Print a test of the hypothesis that the autocorrelation is zero (see Section 9.6.2).

Use the function from the previous questions on the formulated model. Also discuss if it is at all reasonale to make the plot of residuals as a function of the order and test the auto-correlation in the presented case.

- f) **Standardized and studentized ressduals (Optional and possibly difficult):** Expand the residual analysis function such that it is possible to chosse between raw, standardized and studentized residuals, see section 9.10.