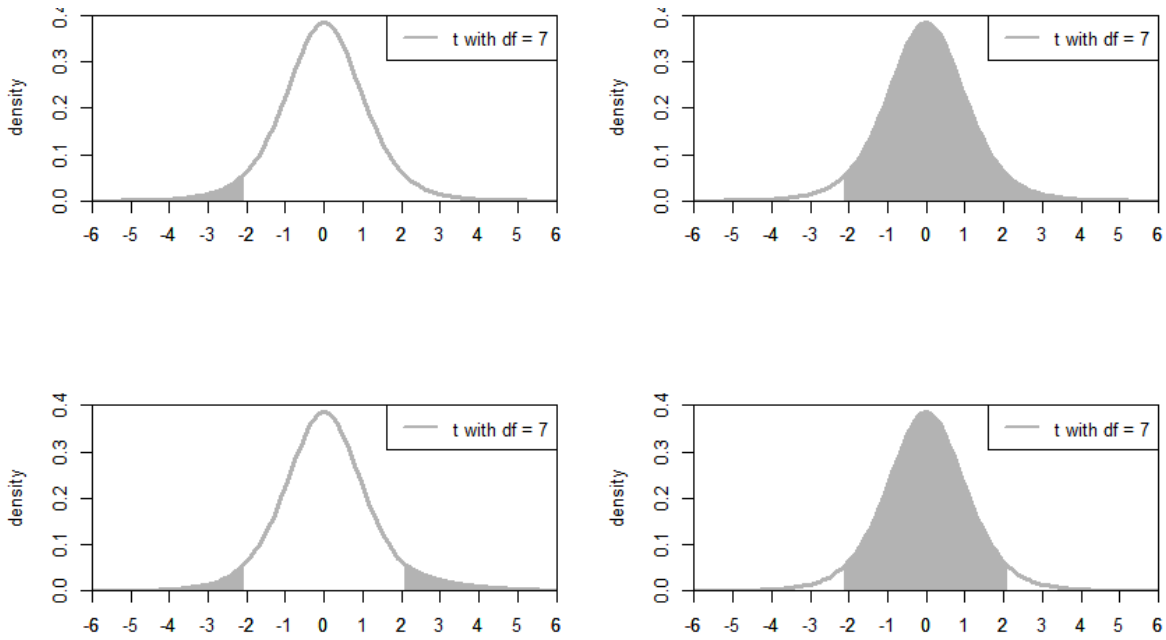
Exercises for exercise class 2 in MMS075, Jan 29, 2020

1. One of the graphs below corresponds to the testing of relationship between response and a single predictor with an associated t-test value of -2.1 in that the shaded area under the curve equals the p-value corresponding to the t-test.



   a) Which one is the corresponding graph?
   b) Explain the meaning of the p-value in words.
   c) Based on this test, can you conclude that there is a relationship between the response and the predictor? (Hint check table about t distribution below!)
   d) What was the number of observations that the model was based on?

Degree of freedom	1	2	3	4	5	10
97.5% quantile	12.71	4.30	3.18	2.78	2.57	2.23

2. Assume the same background story and data as described in Exercise 1 in Exercise class 1: A (hypothetical) very large company called Maintain-IT is responsible for a project task that needs to be repeated every year. They want to determine how the number of employees assigned to the project affects the completion time. An analyst at Maintain-IT decides to use simple linear regression to model this dependence, based on the following observations:

Year	Employees in project	Completion time (days)
1	70	20
2	30	60
3	10	100
4	90	20

We have computed that the least square coefficients are -1 for the slope and 100 as the intercept for the resulting model. Using this model answer the following questions:

a) Having 44 employees assigned to the project, what would be the expected completion time?
b) How many employees need to be assigned to the project for a completion time of approximately one month?
c) If there is exactly one month available for completion with a strict deadline at the end of that period, does this change the number of employees that should be assigned to the project?

3. Consider the multiple linear regression model with sales (in 1000 units) as response and the usual 3 predictor variables of TV advertisements, radio advertisements and newspaper advertisements as predictors. The R output of this model is given below.

```
Call:
lm(formula = AdData$sales ~ AdData$TV + AdData$radio + AdData$newspaper)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.938889   0.311908   9.422   <2e-16 ***
AdData$TV       0.045765   0.001395  32.809   <2e-16 ***
AdData$radio    0.188530   0.008611  21.893   <2e-16 ***
AdData$newspaper -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

a) Is there a relationship between at least one type of advertisement and sales? Where do you see this in the output above?
b) Formulate the interpretation of the following quantities:
   - 0.86 in the row of AdData$newspaper
   - <2e-16 in the row of AdData$radio
   - <2e-16 in the row of (Intercept)
c) Specify an approximate confidence interval for each coefficient.
d) The management decides to spend 115000$ on TV advertisements and 40000$ on radio advertisements. How many sold units does the multiple linear regression model predict for these values?
e) The management sets selling 5000 units as a target. Suggest different ways of achieving this using the above model and specify the associated costs.

4. Group discussion: come up with at least one example related to logistics where simple linear regression could be used and at least one example related to logistics where multiple linear regression may be relevant.

5. Prove that in case of linear regression, $0 \leq R^2 \leq 1$. (Hint: using the formula for computing R^2, show that R^2 is always at least 0 and then that R^2 is always at most 1.)

6. An analyst has used a multiple linear regression model to predict the sales of products in development using the novelty value of the product, its relevance to the market (both on a scale of 0-100 with large values corresponding to more novel and more relevant products) and the advertisement costs in 1000$. However, under serious time pressure while preparing a report summarizing the results, the analyst forgets to copy-paste all relevant information to the report. The resulting table looks like this:

Parameters	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	37.7015	0.798	0.432
Novelty	0.3469	5.139	2.33e-05***
Relevance	0.3997	21.646	< 2e-16***
Advertisements	0.3782	16.277	3.75e-15***

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53.08 on 26 degrees of freedom
Multiple R-squared:  0.9648, Adjusted R-squared:  0.9607
F-statistic: 237.5 on 3 and 26 DF,  p-value: < 2.2e-16
```

Fed up with the work conditions, the analyst resigns soon after preparing the report and starts a trip around the world without leaving any contact information. However, the management immediately needs to decide the advertisement strategy for three new products, product 'A' with Novelty = 90, Relevance = 20, product 'B' with Novelty = 30, Relevance = 40 and product 'C' with Novelty = 70, Relevance = 80.

a) Does it make sense to use the multiple linear regression model to decide the advertisement budget?
b) Can we specify the estimates and approximate confidence intervals for each predictor?
c) At what advertisement budget could we expect to sell 1000 units of each product?

Answer as many questions as you can, based on the presented information!

7. Feedback quiz (optional): Go to www.menti.com and use the code 77 60 44.