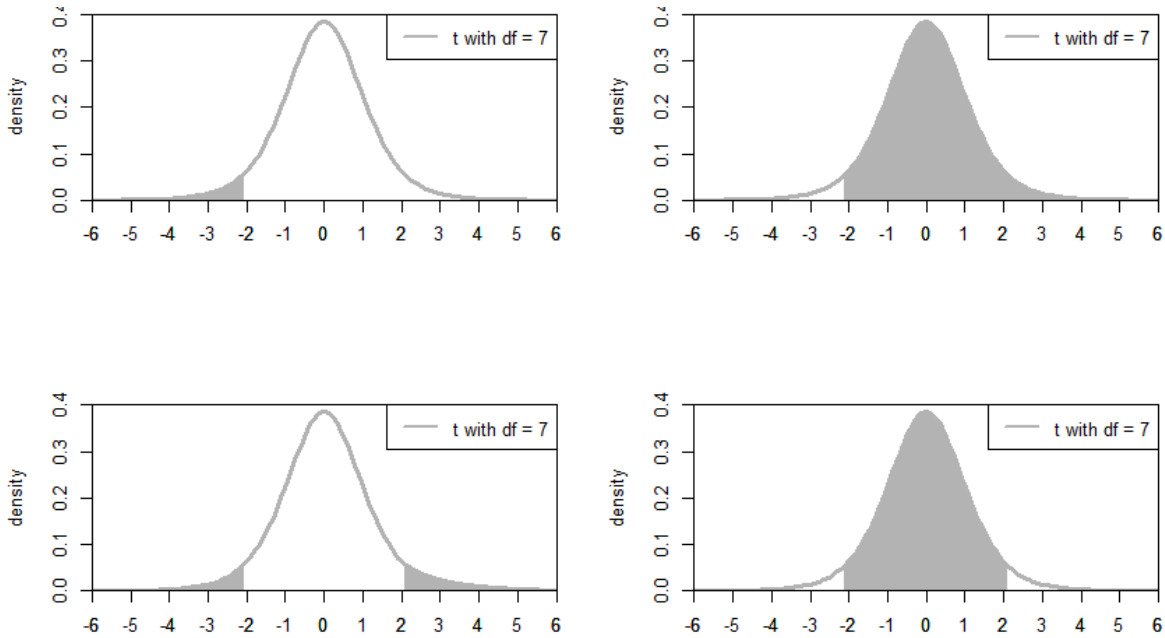
Exercises for exercise class 2 in MMS075, Jan 29, 2020

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.
   Disclaimer: this exercise turned out to be more difficult and more theoretical than I intended it to be. Read the solution below carefully and try to understand it, because it improves your understanding of the t-test; however, it is unlikely that the exam would include such a theoretical question.



   a) Which one is the corresponding graph?
   The p-value is the probability of having a value that is at least as extreme as the test statistic value, assuming that the null hypothesis holds. For testing the relationship between response and a single predictor, the t-statistic has t distribution with n-2 degrees of freedom under the null hypothesis; this distribution is centered at 0 and its extreme values are those that are far away from 0, both in the positive and the negative direction. Therefore, the graph corresponding to the p-value is the one in the lower-left corner, where the area for values smaller than or equal to -2.1 and values greater than or equal to +2.1 are shaded.

   b) Explain the meaning of the p-value in words.
   See the first sentence of the answer to part a) for the interpretation of the p-value in general. In case of testing the relationship between response and a single predictor, the null hypothesis is that there is no relationship. Therefore, the p-value here is the probability that the test statistic is below -2.1 or above +2.1, or in other words, its absolute value is at least 2.1:

$$\Pr\left(\left|\frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}\right| \geq 2.1\right)$$

   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!)
   We could conclude that there is a relationship between response and the predictor if we could reject the null hypothesis of no relationship. We could reject the null if

the value of the test statistic is extreme enough, i.e. if its absolute value is above the critical value specified in the table below. The critical value for the t distribution with df=7 is between the critical values for df=5 and df=10. The absolute value of the test statistic is 2.1, and that is not larger than either of these critical values. Therefore, the conclusion is that the test statistic is not extreme enough to reject the null hypothesis, hence we CANNOT REJECT that there is no relationship. In other words, the t-test DID NOT PROVIDE SUFFICIENT EVIDENCE OF A RELATIONSHIP between the predictor and the response.

d) What was the number of observations that the model was based on?
In case of testing the relationship for simple linear regression, df=n-2. The graphs show that df=7. Therefore, basic algebra gives that n=9, i.e. there were 9 observations.

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?
Denoting the number of employees by x and the completion time by y, the linear model gives us an estimate of y for each value of x:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Since we know that the intercept is 100 and the slope is -1, we can plug in the value of x for which we want to make a prediction, i.e. x=44, in this equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 100 - 1 \cdot 44 = 56.$$

This gives that in case of 44 employees assigned to the project, the expected completion time is 56 days.

b) How many employees need to be assigned to the project for a completion time of approximately one month?
We again need to use the equation of

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

But in this case, we know the y-hat value and are looking for x. Approximately one month can be interpreted as 30 days, so this is the value we plug in:

$$30 = 100 - 1 \cdot x$$

Solving this equation using basic algebra gives that x=70. The conclusion is therefore that 70 employees need to be assigned to the project to get a completion time around 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?

Yes, it does. Having 70 employees gives a predicted completion time of 30 days, but it can easily happen that the actual completion time will be a bit less or a bit more (we will discuss at the next lecture how much more), and if the deadline is strict, then having a longer completion time is not acceptable. It is likely that many more employees would be needed to be sufficiently sure about meeting the strict deadline.

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?

Yes, there is. When testing this, the null and alternative hypotheses are that

$$H_0 : \beta_1 = \beta_2 = \ldots \beta_p = 0$$

$$H_a : \text{at least one } \beta_j \text{ is not zero}$$

and an F-statistic is computed to test this. The last line of the R output shows the value of the F-statistic and that the associated p-value is extremely small (p-value: < 2.2e-16). Therefore, this test gives strong evidence of a relationship between at least one type of advertisement and sales.

b) Formulate the interpretation of the following quantities:

- 0.86 in the row of AdData$newspaper

Interpretation: this is a very large p-value, indicating that if we have radio and TV in the model, then adding newspaper does not help in predicting sales. In other words, in the presence of TV and radio, newspaper is not a significant predictor.

- <2e-16 in the row of AdData$radio
  Interpretation: this is a very small p-value, indicating that even if we have TV and newspaper in the model, then adding radio significantly contributes to predicting sales. In other words, even in the presence of TV and newspaper, it is important to add radio to the model.
  <2e-16 in the row of (Intercept)
  Interpretation: this is a very small p-value, indicating that the intercept term is significantly different from 0. This provides strong evidence that even without any TV, radio and newspaper advertisements, sales will not be zero (but rather some positive number, considering also that the estimated intercept is positive).

c) Specify an approximate confidence interval for each coefficient.
It was discussed during the lecture that in case of n≥30, one can use the formula of estimate ± 2*SE(estimate) to determine an approximate confidence interval. Here we can see in the third line from below in the R output that the degrees of freedom in the model is 196. This indicates that the sample size is n=200 (because df=n-p-1, so n=df+p+1) and we can feel free to use this simple formula. This means that the confidence interval…
… for the intercept is 2.938889 ± 2*0.311908; that is: we are 95% confident that 2.315 ≤ intercept ≤ 3.563;
… for the TV coefficient is 0.045765 ± 2*0.001395, that is: we are 95% confident that 0.043 ≤ TV coefficient ≤ 0.049;
… for the radio coefficient is 0.18853 ± 2*0.008611 that is: we are 95% confident that 0.171 ≤ radio coefficient ≤ 0.206
… for the newspaper coefficient is -0.001037 ± 2*0.005871, that is: we are 95% confident that -0.013 ≤ newspaper coefficient ≤0.011.
Note that all confidence interval for newspaper coefficient contains 0 while the other confidence intervals do not contain 0. This is another way of seeing that radio and TV are significant predictors even in the presence of all others, but newspaper is not significant in the presence of TV and radio, and that the intercept is significantly different from 0.

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?
The text does not mention any spending on newspaper ads, so let us assume that the management spends 0 on newspaper advertisements. Knowing the amounts spent on each advertisement form, we can define values for the three predictors of the model: since we were counting in 1000$ spendings, the relevant values are: TV=115, radio=40, newspaper=0. We can use the formula for predicting in the multiple linear model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

In this case, we know the estimated coefficients (i.e. the beta-hat values) from the R output, which gives the following formula for the predicted sales:

$$\widehat{\text{sales}} = 2.939 + 0.046 \times \text{TV} + 0.189 \times \text{radio} - 0.001 \times \text{newspaper}$$

Plugging the values of TV=115, radio=40, newspaper=0 in this equation gives that:

$$\widehat{\text{sales}} = 15.789$$

The sales variable indicates the number of products sold in 1,000 units, so having an estimate close to 16 for this variable means that we predict about 16,000 units (more precisely: 15789 units) to be sold when the management spends 115000$ on TV advertisements and 40000$ on radio advertisements.

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.

There are many ways to do this. Two were discussed in class:

- Consider TV advertisements only, i.e. set radio=0 and newspaper=0. In that case, the formula for predicting reduces to a simpler form:

  $$\widehat{\text{sales}} = 2.939 + 0.046 \times \text{TV}$$

  We want to reach 5000 sold units – as the sales variable indicates the number of products sold in 1,000 units, we need the TV value that gives a predicted value of 5 for sales. Therefore, we need to solve the following equation:

  $$5 = 2.939 + 0.046 \times \text{TV}$$

  This gives a value of TV=44.804. We should again remember that the value of the TV variable corresponds to the TV advertisement costs in 1000$. Therefore, having a value of TV around 44.8 corresponds to a cost of 44800$. The conclusion is that spending 44800$ on TV advertisements and zero on other advertisement forms yields predicted sales of 5000 units.
- Consider radio advertisements only, i.e. set TV=0 and newspaper=0. In that case, the formula for predicting reduces to a simpler form:

  $$\widehat{\text{sales}} = 2.939 + 0.189 \times \text{radio}$$

  and we need to solve the following equation:

  $$5 = 2.939 + 0.189 \times \text{radio}$$

  This gives a value of radio=10.905, which corresponds to a cost of 10905$. The conclusion is that spending 10905$ on radio advertisements and zero on other advertisement forms yields predicted sales of 5000 units.

The latter cost is much less than the cost required when considering TV ads only, so we conclude that considering radio advertisements only is a much better strategy. This is not surprising – the coefficient of radio is much larger than the coefficient of TV, so every additional $1000 spent on radio advertisement results in a much steeper increase in sales than spending the same amount on TV advertisements, so it requires less money to reach the desired 5000 sold units this way.

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.

   Examples when simple linear regression could be relevant:

   - Model the number of containers handled in a harbor based on the number of cranes;
   - Model the time or cost associated with transporting goods based on the distance between the source and destination.

Multiple linear regression could give a refinement of the models; one example is that the predicted time or cost associated with transporting goods could also depend on whether the destination is inside or outside the EU. This could be represented with a categorical variable; including such variables in linear regression models will be discussed at the next lecture.

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.)

First, we show that $R^2 <= 1$. As squared numbers are always nonnegative and RSS and TSS are sums of squared numbers, we have that RSS ≥ 0, TSS ≥ 0 and RSS/TSS ≥ 0. Therefore, we have that:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \leq 1$$

Now we prove $R^2 >= 0$ using the same formula. We note that

$$1 - \frac{\text{RSS}}{\text{TSS}} \geq 0 \text{ if and only if } \text{TSS} \geq \text{RSS}$$

The latter inequality is easiest shown using the geometrical interpretation: the definition of the least squares line is that its RSS is smallest among the sum of squares among all possible lines. Since TSS can be seen as a sum of squares relative to the line with slope 0 and intercept $\bar{y}$ (see slide 10 in Lecture 2), the corresponding sum of squares cannot be smaller than RSS.

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!

This exercise will be considered again in Exercise class 3. Please think about it and try to solve it before that class.

Hint: think about the way that t-values are computed from coefficient estimates and standard errors. The corresponding formula will allow you to reconstruct the estimated coefficients from the t-value and standard error columns. Once you have those available, this exercise will be similar to exercise 3 discussed above.

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