# Tasks chapter3: Linear regression

For the “location” assigned to your dyad, use the SYNOP observations and forecast data from ECMWF with a forecast horizon of 36 hours contained in `location_obs_ECMWF_2009-2022.rds`. The file contains several forecast parameters for which acronyms, description and units are given in `ECMWF_selected_parameters.pdf`. The observations are `rr6hObs`, the precipitation sum over the previous 6 hours (mm), `t2mObs`, the temperature 2 m above ground (Celsius), and `tdObs`, the dew point temperature (Celsius).

The R-lab session at the end of chapter 3 in the ISL book will show you all necessary R commands.

0. Read in the data with `readRDS()` as in the previous task and assign it to your `location`. Also convert `t2m` from Kelvin to Celsius
1. Use the `lm()` function for a simple linear regression with `location$t2mObs` as the response and `location$t2m` as the predictor and assign it to the variable `lm.fit`. R has a formula notation with the “~” sign: `lm(t2mObs ~ t2m, data = ibk)`. The `data = ibk` saves you from having to type `ibk$t2mObs`.
2. What are the coefficients (intercept and slope) of the linear regression? Use `print(coef(lm.fit))`
3. Use `summary(lm.fit)` to print the results.
4. Is there a relationship between the predictor and the response and how strong is it?
5. How much of the variance is explained?
6. What does the value of the F-statistic tell you about the possibility of rejecting the null-hypothesis that the slope $\hat{\beta}_1$ is (close to) zero?
7. How large is the 95% confidence interval for the intercept and slope of the linear regression? By how many percent does the slope change from the 2.5% level to the 97.5% level of the confidence interval? By how many Kelvin the intercept? Confidence levels are contained in `confint(lm.fit)`
8. Make a scatterplot of `(location$t2m, location$t2mObs)` with small symbols (option of `cex=0.1`) and add a red line showing the linear regression (`abline(lm.fit, col = "red")`). Add the pivot point (= mean of (x), mean of (y)) using `points()` with ‘col = “orange”.
9. Which data point has the highest leverage? What does that mean? How much higher is it than the average leverage (cf. p99 of ISL book)? Use `hatvalues(lm.fit)` for the h-statistic and `which.max()` to get the index of the observation with the highest h-statistic (== maximum hatvalue). What temperature was that?
10. Is there any outlier? (what is the difference between outlier and high leverage point?). Use a plot of studentized residuals (almost all values should be between +-3) produced with `plot(predict(lm.fit), rstudent(lm.fit))`
11. Produce a 4-panel diagnostic plot of the linear regression - as shown in class. Tell R to produce a 2x2 graphics with `par(mfrow = c(2,2))` and then simply do `plot(lm.fit)`. (A remark: if you later want only 1 figure per plot reset with `par(mfrow = c(1,1))`.)
    These plots can be used to spot whether basic assumptions of least-squares regression are violated. Assumptions are that the errors (residuals) have a Gaussian distribution, that they are centered on the regression line and that their variance does not change as a function of x (“homoscedasticity”.
    (Note that in the plots some (problematic) data points are labeled. Since we are using date (YYYY-MM-DD hh:mm:ss) as index that looks messy.)