Tasks chapter3: Linear regression

David Kurz, Malte Hildebrandt
2023S707737 VU Geostatistik
Universität Innsbruck

March 25, 2023

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.

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 "$\sim$" sign: lm(t2mObs $\sim$ 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))

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{1}$$

Table 1: Coefficients of linear regression

intercept ($\hat{\beta}_0$)	**slope** ($\hat{\beta}_1$)
-1.669422	1.037230

3. Use summary(lm.fit) to print the results.

Table 2: Residuals

min	**1Q**	**median**	**3Q**	**max**
-9.9010	-0.9523	0.2410	1.2517	7.3054

The standard error of an estimator reflects how it varies under repeated sampling.

Table 3: Coefficients

	coefficient	**std.-error**	**t-values**	**p-value**
intercept	-1.669422	0.051533	-32.4	<2e-16
t2m_celsius	1.037230	0.003078	337.0	<2e-16

Residual standard error: 1.907 on 5043 degrees of freedom (52 observations deleted due to missingness).

Multiple R-squared: 0.9575, Adjusted R-squared: 0.9575

---

4. Is there a relationship between the predictor and the response and how strong is it?

   R-squared = 0.9575, so the relationship between predictors and response is very strong positive!!

---

5. How much of the variance is explained?

   If the R-squared values is approaching to 1, it means that our predictors explain much of our data variability (much of our variance).

---

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?

   F-statistic: 1.136e+05 on 1 and 5043 DF, p-value: < 2.2e-16
   The F-statistic quantifies the difference in error between the model, assuming your null hypothesis (i.e. $\hat{\beta}_1 = 0$) and your fitted model. It simply a formal way of determining if a coefficient belongs to the model. If the p-value recieved from the F-testing statistics is small then we reject the null (because the probability of occurrence is so small), and say $\hat{\beta}_1 = 0$ here. So assuming $\hat{\beta}_1 = 0$ and rejecting the null hypothesis here must have to something that we expect the p-value small.

---

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, level = 0.95).

Table 4: Confidence intervalls

	2.5%	**97.5%**	**change**
intercept ($\hat{\beta}_0$)	-1.770450 °C	-1.568394 °C	0.202056 K
t2m_celsius ($\hat{\beta}_1$)	1.031196	1.043264	101.2%

---

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".

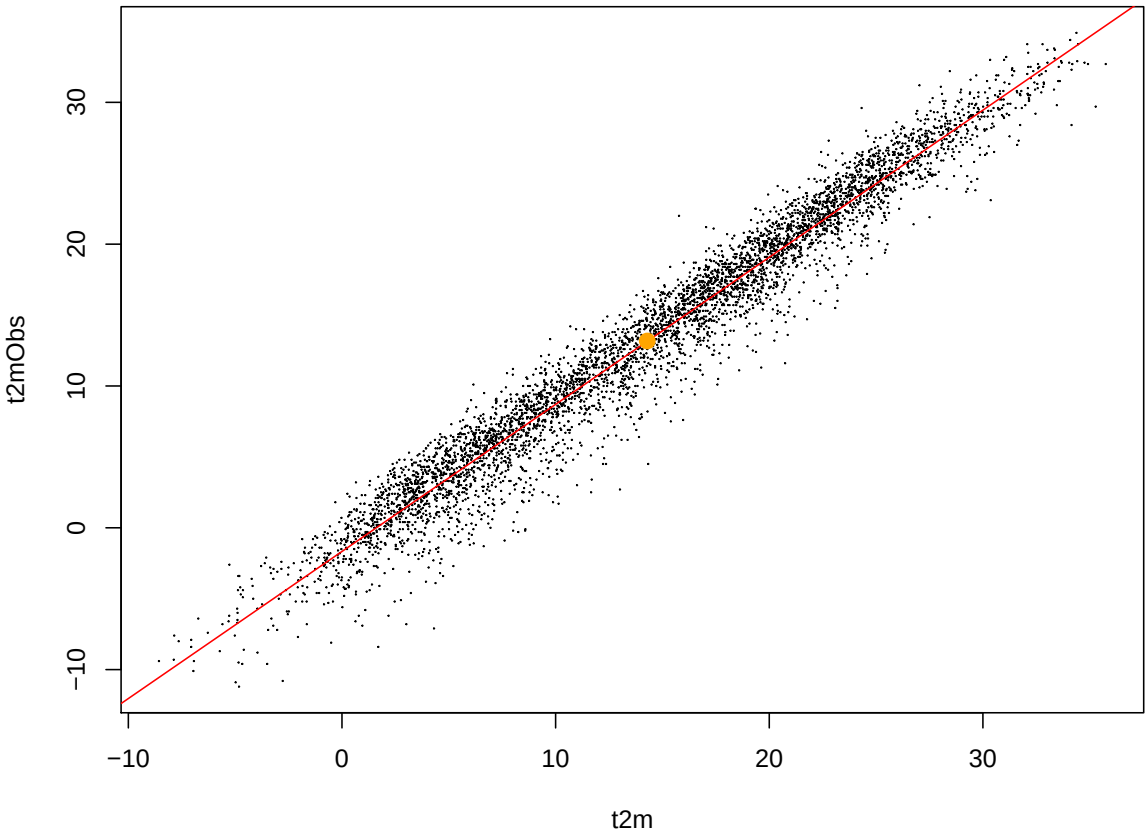


Figure 1: scatterplot with linear regression (red line)

---

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?

   - **max:** 0.00156 (data point 1129) ⇒ data point with high leverage has a large potential to affect the overall regression analysis.
   - **mean:** 0.000396
   - **difference:** 0.00116
   - **temperature at max:** -8.4°C (Obs.) -7.05°C (Pred.)
     There is no change in the predicted values, although the observations change considerably at this point.

---

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))

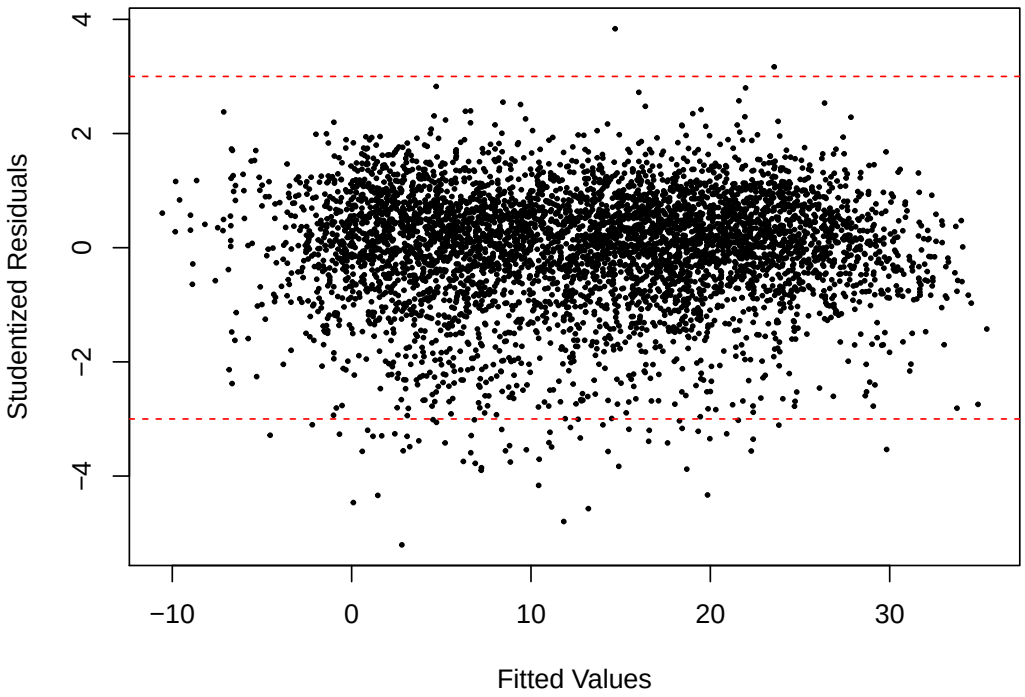


Figure 2: Residuals for identifying outliers

- Is there any outlier?
  Many negative, only two over +3
- What is the difference between outlier and high leverage point?
  A high leverage point is a data point that has a relatively extreme value for one or more predictor variables. High leverage points can have a large influence on the position of the regression line, but they may not necessarily be outliers. In some cases, high leverage points may be legitimate observations that have a unique combination of predictor variable values.

---

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.)

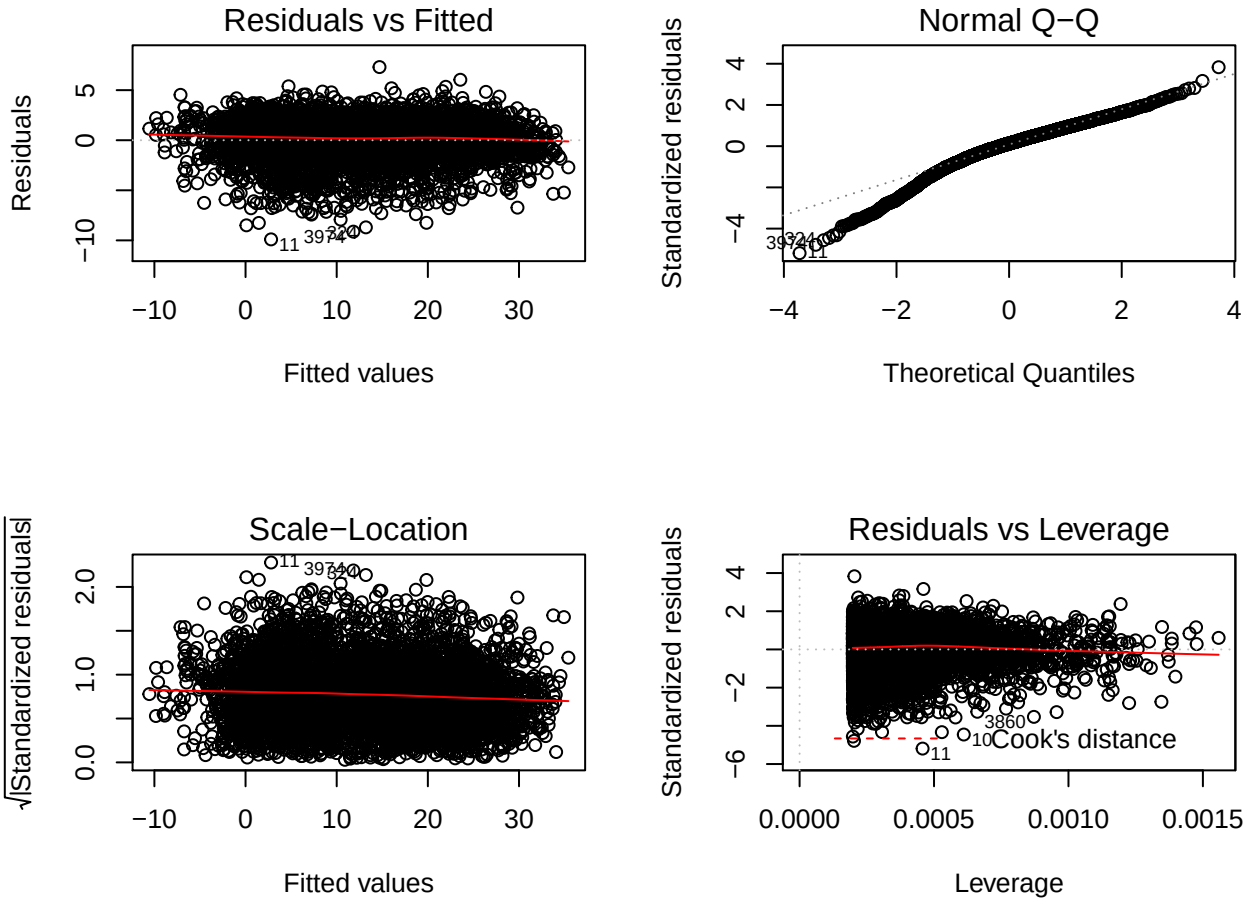
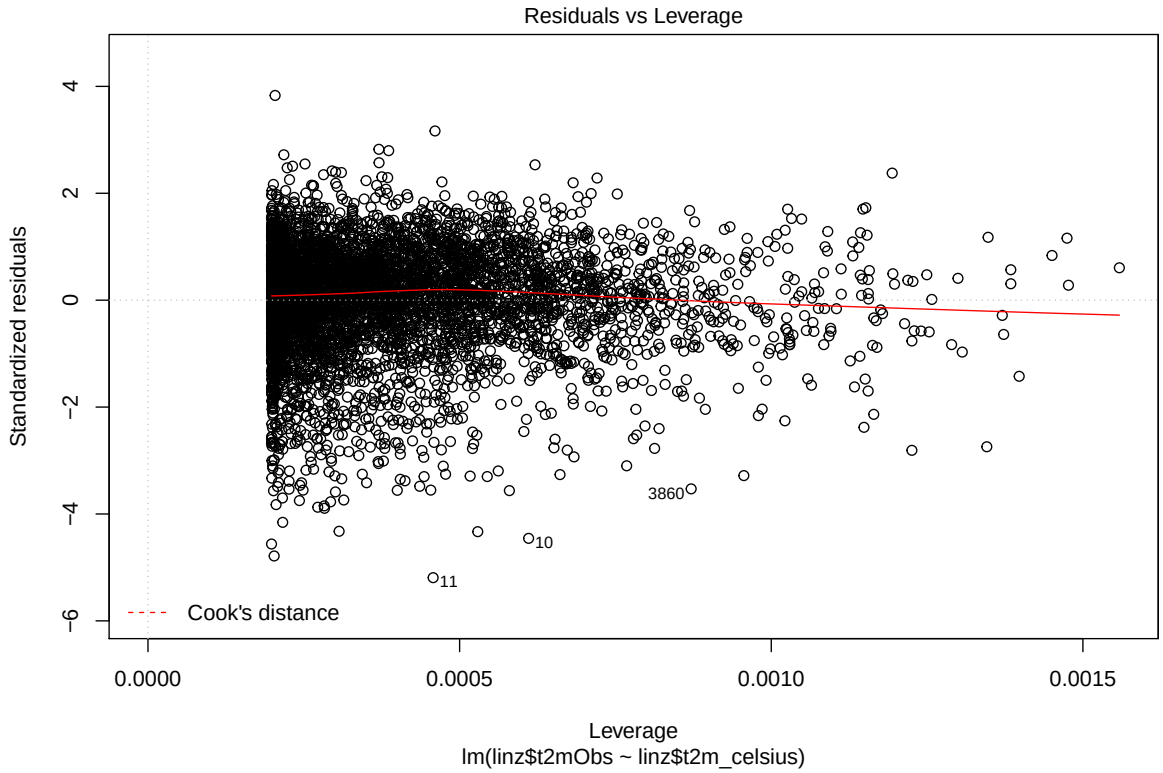


Figure 3: diagnostics of linear regression



No point falls out of cooks distance. This means there aren't any overly influential points in our dataset. So no basic assumptions are violated.