Solution Sheet 11

1. a) From the scatterplot, it is evident that the first 7 points are described very well by a straight line. The last measurement, however, is totally out of line.

Although the outlier may just be a huge mistake, it is also posible that the linear model is simply not good enough for describing the relationship between depth and temperature. (For instance, the relationship might be quadratic or piecewise linear. Further explanations can also be considered, such as hot springs, etc.)

b) Let X be the depth, and Y the temperature. Their empirical correlation is

$$\rho_{\mathrm{X,Y}} = \frac{\sum_{i=1}^{7}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{7}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{7}(y_i - \overline{y})^2}} = -0.99 \ ,$$

where $\overline{x}$ and $\overline{y}$ must also be computed without the outlier.
When the outlier is excluded, the depth and temperature exhibit very strong negative correlation; when the outlier is included, however, their correlation is positive (0.6)!

c) In the above plot, both regression lines are drawn exactly; the effect of the outlier is to rotate the regression line by nearly 90°. In other words: least-squares regression is highly sensitive to outliers!
Without the outlier, the coefficient estimates are as follows:

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{(6 + 1.81)(0 - 0.6) + \ldots + (-8.9 + 1.81)(1.2 - 0.6)}{(0 - 0.6)^2 + \ldots + (1.2 - 0.6)^2} \\ &= -13.64\end{aligned}$$

and

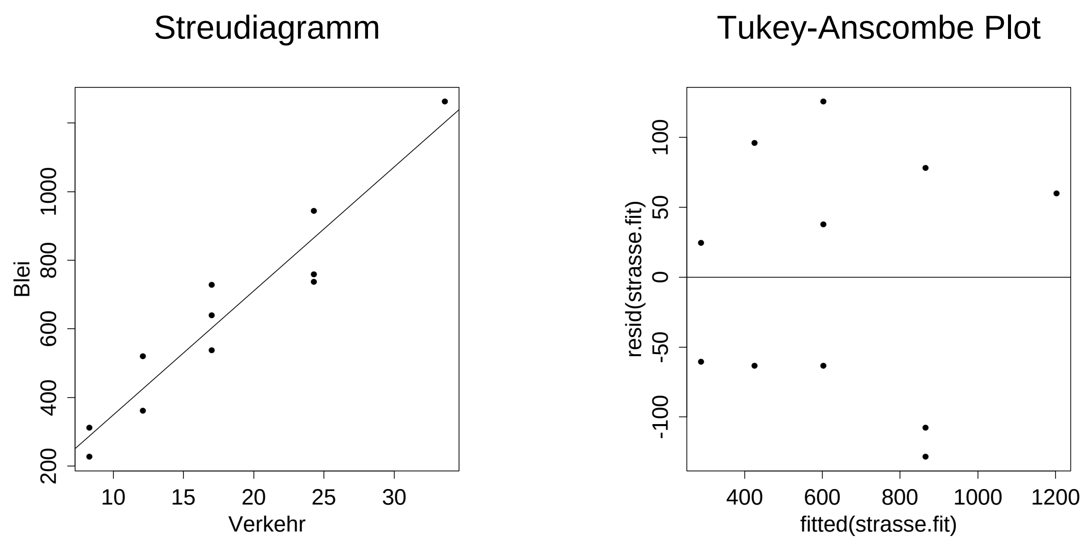
$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} = -1.81 - (-13.64) * 0.6 = 6.37.$$

Including the outlier gives us $\widehat{\beta}_1 = 2.47$ and $\widehat{\beta}_0 = -2.86$ instead.
These estimates can also be seen in the output of **R**.

d) Without the outlier:
Null hypothesis H_0: $\beta_1 = 0$;   alternative hypothesis H_A: $\beta_1 \neq 0$.
Test statistic: $T = \frac{\widehat{\beta}_1 - 0}{\widehat{\sigma}(\widehat{\beta}_1)} = \frac{-13.64}{1.01} = -13.5$.
Rejection set at 5%: $\mathcal{K} = \{T : |T| > t_{7-2,\,0.975}\} = \{T : |T| > 2.57\}$
Outcome: H_0 is rejected; the slope cannot be 0.

Including the outlier gives us $\widehat{\beta}_1 = 2.47$ und $\widehat{\sigma}(\widehat{\beta}_1) = 1.33$. This yields the p-value 0.112, which does not lead to the rejection of H_0.
These outputs are obtained in **R** using the command `summary`. Thus for example:

2. a)



b) *> summary(highway.fit)*

```
Call:
lm(formula = blei ~ verkehr, data = highway)

Residuals:
    Min      1Q  Median      3Q     Max
-128.43  -63.13   24.52   69.32  125.72

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -12.842     72.143  -0.178    0.863
verkehr       36.184      3.693   9.798 4.24e-06 ***
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1

Residual standard error: 92.19 on 9 degrees of freedom
Multiple R-squared: 0.9143,     Adjusted R-squared: 0.9048
F-statistic: 96.01 on 1 and 9 DF,  p-value: 4.239e-06
```

The estimates we obtain are $\widehat{\beta}_0 = -12.842$, $\widehat{\beta}_1 = 36.184$ and $\widehat{\sigma}_e = 92.19$ (residual standard error).

$$\Longrightarrow Lead = -12.842 + 36.184 \cdot Traffic$$

c) **t-test** for β_1: null hypothesis H_0: $\beta_1 = 0$, alternative hypothesis H_A: $\beta_1 \neq 0$

Test statistic: $T = (\widehat{\beta}_1 - 0)/\widehat{\sigma}(\widehat{\beta}_1)$. Under H_0, we have $T \sim t_{n-2}$, and thus $T \sim t_9$ here

Critical set: the tabulated value is $t_{9,0.975} = 2.26$ and thus $\mathcal{K} = \{|T| > 2.26\}$.

Value of the test statistic: $T = 36.184/3.693 = 9.8 > 2.26$ (cf. computer output). The slope β_1 is very significantly different from zero (`p-value` $= 4.24e - 06$).

d) $x = 40 : y = -12.842 + 36.184 \cdot 40 = 1434.518$

3. 1) a
2) a
3) b
4) d
5) a
6) c
7) c