Exercise 5: Interpreting linear regression

We shall in this exercise consider data from an experiment where the objective was to relate the tissue strength of paper produced in a factory to three variables characterizing the production process:

- mixture of two mass types (x_1)
- temperature (x_2)
- pressure (x_3)

The experiment was performed by varying the levels these variables over appropriate regions and measuring the resulting tissue strength (y).

The model connecting the response to the covariates is at the outset assumed to be a second degree polynomial

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_4 x_2^2 + \beta_6 x_3^2 + \beta_7 x_1 x_2 + \beta_8 x_1 x_3 + \beta_9 x_2 x_3 + \epsilon$$

where the mathematical notation is as in Lecture 3. The table shows the least squares estimates $(\hat{\beta}_j,\ j = 1, 2, \ldots, 10)$ of the ten regression coefficients along with their estimated standard deviations based on on data from a real experiment at a Norwegian research institution.

	$\hat{\beta}_j$	stan$(\hat{\beta}_j)$	r_j^2
Intercept	4.650	0.90	
x_1 (mass type)	0.121	0.041	0.15
x_2 (temperature)	0.033	0.043	0.01
x_3 (pressure)	−0.134	0.046	0.18
x_1^2	−0.032	0.046	0.01
x_2^2	−0.065	0.046	0.02
x_3^2	0.073	0.043	0.03
$x_1 x_2$	−0.028	0.043	0.01
$x_1 x_3$	0.138	0.043	0.18
$x_2 x_3$	0.105	0.047	0.09

Table 1: Estimated regression coefficients with standard deviations. The squared correlations between the response and the covariates are give to the far right.

Also given (in the column to the far right) are the estimated correlation coefficients (r_j^2) between the response y and the explanatory variables x_1, x_2, x_1^2 and so on.

The levels of the mass type, temperature and pressure were under control of the experimenters and carefully selected to make the experiment *orthogonal*. This means that the correlations between all nine x-terms on the right of the intercept in the model equation had correlation 0. We argued in Lecture 3 that this produced estimates of the regression coefficients that were as accurate as possible. It has also the convenient consequence that the multiple correlation coefficient R^2 can be decomposed into separate contributions from the nine x-terms, i.e.,

$$\mathrm{R}^2 = r_1^2 + r_2^2 + \cdots + r_9^2$$

so that the squares of the correlations represent the impact on term j on the variation of the response y.

The significance of the regression coefficients can be assessed by computing their *t-ratios*[5], that is their estimated values divided on their estimated standard deviations. Use the t-ratios along with the squared correlations to suggest a predictor for tissue strength based on mass type, temperature and pressure. Only include terms which will improve the prediction according to your judgement. Can you see arguments against using all the 10 terms in the basic modelling equation?

[5]The number of observations of the experiment was 29, which implies that the number of degrees of freedom of the t-statistics are $29 - 9 - 1 = 19$.