Exercises 3 solutions

1. (a) **First: assessing relations between y and expl. variables — are there any? linear? non–linear?** In looking at the dependence of Y on x_1, x_2 and x_3, there does not appear to be any strong relationship between Y and x_1 or between Y and x_3. Only three different values of x_2 were used (two of which close together and one further apart — these will have more "leverage" than the others) and the scatter at each of these points at each value is small enough to suggest a linear regression model fitted to these data will need to involve a nonmonotonic (for example quadratic) term in x_2.

   Secondly: assessing relations among expl. variables. Although the trend is for x_3 to (roughly linearly) increase with x_1 there does not appear to be any strong linear dependencies between the three explanatory variables.

   (b) Model 1 has $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$ for $i = 1, \ldots, 27$.

   Additional statistical assumptions are that the errors $e_1, \ldots, e_{27}$ are independent $\mathcal{N}(0, \sigma^2)$.

   NOTE: the distributional assumption (either about the errors or about the responses) is part of the model! Also, if you index Y_i by i then the explanatory variables and error term must be indexed by i.

   (c) Under the above assumptions, the F-test shows that there is strong evidence that the response Y depends on at least one of the explanatory variables, i.e., at least one of the radiation, the soil moisture tension and the air temperature is informative about the amount of vitamin B_2 in turnip green.

   (d) The percentage of the total variation explained by the regression model, R^2, is reasonably high at 75.5%, which appears to provide some support for the model – at least a large amount of variability in Y can be explained. The scatter in the residuals v fitted plot does not look to be too bad. There are two clear groups along the x-axis, which is not a violation of the model assumptions, and the only reason for concern would be that the two clearly smallest and the largest residual appear all in the group with larger fitted values (though this can be due to random variations because there are more points in that group anyway).

   The two largest negative residuals, by looking at the normal plot, can be seen to have standardised values less than -2. This is one more standardised residual of this size than we would expect in a sample of size 27, assuming the errors are normally distributed, though the plot still looks not too far away from a straight line.

   However, a clear violation of the model assumptions appears in the plots of the residuals against each explanatory variable under model 1: while the plots for residuals v x_1 and x_3 look fairly random (though possibly with higher variance in the middle), the one for residuals v x_2 has a nonmonotonic (possibly quadratic) appearance (as far as one can say from three values of x_2) reflecting that for Y v x_2 in the matrix plot. The latter plots suggest the inclusion of a quadratic term in x_2 in the model, ie as in model 2, because this is the simplest possible extension to nonlinearity that allows to fit functions with a maximum or minimum inside the range of the x-variable. (Hence a reason why we need not bother looking at the remaining output for model 1.)

   Note: interpretation of plots is to some extent subjective; you may see things differently. Any sensible comments are acceptable. However, the nonlinearity in residuals v x_2 should definitely be spotted.

   The **residual versus explanatory variables plots are often most informative** in terms of how the model should be specified, so make sure you comment on those.

   (e) For model 2, the residual standard error has decreased to 6.104 and R^2 has noticeably increased to 91.0%; the plot of residuals v fitted, although still showing two groups, has much more scatter to it than for model 1; also the strong quadratic trend to the plot of residuals v x_2 has gone! There are now three standardised residuals that look to be greater than 2 in magnitude, one or two more than expected. Despite this, model 2 looks to be a substantial improvement in fit compared to model 1. In the following two parts, it will be assumed that the assumption of normally distributed errors is not unreasonable: the normal plot is fairly straight except for a little deviation from those observations with the largest standardised residuals.

(f) The t-test for the coefficient of x_2^2 has a very small P-value, which means that inclusion of the quadratic term improves the explanation of vitamin B_2 significantly *given the other terms in the model.*

(g) The t-test for the coefficient of x_1 gives a P-value of 0.18, which is not very small. Depending on whether there is a practical use for removing variables, this term could be removed from the model *given the inclusion of the other terms in the model.* A proper interpretation would be that there is no evidence that radiation contributes any information to the explanation of vitamin B_2 in turnip green *that is not also in the remaining variables.*

(h) Removing x_1 from model 2 gives model 3 whose fitted form is

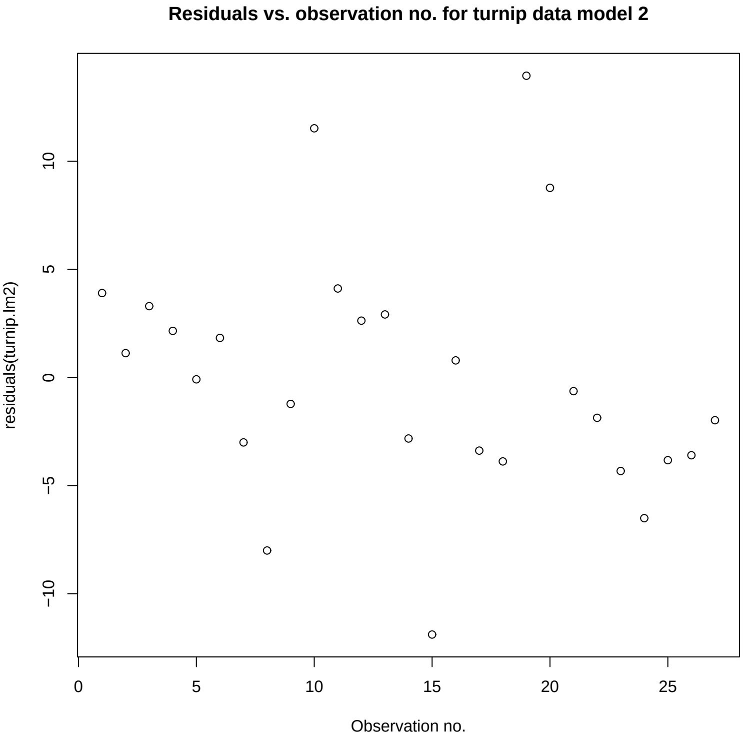
$$\hat{y} = 120.627 + 490.414x_2 - 5.716x_3 - 1107.853x_2^2.$$

(Note that the fitted model is not usually regarded as random, and does not include an error term — it stands for the 'fitted line' (or similar in higher dimensions) and does not inform about the error terms).
The residual standard error of 6.223 has only increased slightly and the value of R^2 has only dropped slightly by the removal of x_1. The residual v fitted plot is not much different from that for model 2 and could be viewed to have sufficient random scatter to support the adequacy of the predictor for model 3. In the normal plot, the points look to be straighter than those for model 2, sufficiently straight to suggest that there is no strong evidence against the assumption of normally distributed errors. However, there are 2 standardised residuals with magnitude just greater than 2; these correspond to observations 15 and 19 which also have the highest Cook's distance.

So model 3 looks OK, but perhaps further investigation could be carried out, eg by seeing what happens if observation 15 or 19 is omitted or if a robust method is applied (cf. Exercises 5). Also, as will be discussed in Section 2.4, it is not always a good idea to remove a non-significant variable, so model 2 still has its merits. (Using the material in Section 2.4, one could compare them using cross-validation.)

Appendix. You were not asked to interpret the estimated regression coefficients in question 1. However, when interpreting the results of a regression analysis in full, this should be done (at least as far as it is interesting for the application). For example, in model 3, when increasing the temperature x_3 by 10 Fahrenheit *and fixing the other variables*, the vitamin B_2 quantity can be expected to go down by 5.716 mg per gram. This is the standard way how to interpret estimated regression parameters. However, if squared terms are involved, it's not so straightforward. So there is no single value for the expected increase in vitamin B_2 if the soil moisture tension goes up by 100 (x_2). A reasonable interpretation could be obtained by inserting $\bar{x}_3$ for x_3 and computing the expected value of Y for all three values of x_2 from the formula above. The differences between the outcomes are the expected changes in Y when replacing one value of x_2 by another, fixing the other variable(s).

Below you find the residuals vs. observation order plot for Model 2, not given in the Exercises. Note that a proper interpretation of this would only be possible if the meaning of the observation order were known. However, the plot actually raises some moderate concern because apart from some exceptions (basically the most outlying points), subsequent residuals seem to be somewhat similar to each other, indicating some positive dependence at least for parts of the experiment, and they also seem to go slightly down on average. (Not sure whether this is clearly different from what could be explained by random variation though; in real life, the next thing to do would be to check whether there is any conceivable explanation for dependence in the experiment).



General: in such questions, try to express yourself precisely and use the technical terms in their correct meaning. E.g. ‘correlation’ is a measure of *linear* dependence between two random variables — it does not apply to non–linear relations and it does not apply to qq–plots (the quantiles are not random variables; it is also inappropriate to say that “the relationship is nonnlinear” if only the qq-plot doesn't look linear). OR: null hypotheses are rejected or *not rejected*, they are *never accepted*. OR: remember that words like ‘affect’ or ‘influence’ suggest a causal interpretation which might not be appropriate; try to use ‘is informative for’, ‘dependence’ and ‘association’ instead. OR: remember to state whether dependencies you see are positive, negative, or some other shape.

2. (a) Observe

$$Z_i = \frac{1}{a}(Y_i - \mathbf{u}_i^T\mathbf{b}),\ u_i = (\mathbf{A}^T)^{-1}\mathbf{x}_i,\ \boldsymbol{\beta}_Z = \frac{1}{a}(\mathbf{A}\boldsymbol{\beta}_Y - \mathbf{b}).$$

By plugging in and multiplying by a, the model for Z_i becomes

$$(Y_i - \mathbf{u}_i^T\mathbf{b}) = \mathbf{x}_i^T\mathbf{A}^{-1}(\mathbf{A}\boldsymbol{\beta}_Y - \mathbf{b}) + ad_i.$$

Define $e_i = ad_i$. Observe $e_i \sim N(0, a^2\sigma_Z^2)$. Add $\mathbf{u}_i^T\mathbf{b}$ and replace $\mathbf{u}_i^T$ by $\mathbf{x}_i^T A^{-1}$ again to get

$$Y_i = \mathbf{x}_i^T\mathbf{A}^{-1}\mathbf{A}\boldsymbol{\beta}_Y + \mathbf{x}_i^T\mathbf{A}^{-1}\mathbf{b} - \mathbf{x}_i^T\mathbf{A}^{-1}\mathbf{b} + e_i,$$

which can be simplified to the standard linear model in Y_i.

(b) With the hint,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_Y &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\\ &= (\mathbf{A}^T\mathbf{U}^T\mathbf{U}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{U}^T(a\mathbf{Z} + \mathbf{U}\mathbf{b})\\ &= \mathbf{A}^{-1}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T(a\mathbf{Z} + \mathbf{U}\mathbf{b})\\ &= a\mathbf{A}^{-1}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{Z} + \mathbf{A}^{-1}\mathbf{b}\\ &= \mathbf{A}^{-1}(a\hat{\boldsymbol{\beta}}_Z + \mathbf{b}).\end{aligned}$$

(c)

$$\begin{aligned}\hat{\sigma}_Y^2 &= \frac{1}{N-p}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_Y)^T(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_Y),\\ \hat{\sigma}_Z^2 &= \frac{1}{N-p}(\mathbf{Z}-\mathbf{U}\hat{\boldsymbol{\beta}}_Z)^T(\mathbf{Z}-\mathbf{U}\hat{\boldsymbol{\beta}}_Z).\end{aligned}$$

The statement now follows from

$$\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_Y = a\mathbf{Z}+\mathbf{U}\mathbf{b}-\mathbf{U}\mathbf{A}\mathbf{A}^{-1}(a\hat{\boldsymbol{\beta}}_Z+\mathbf{b}) = a(\mathbf{Z}-\mathbf{U}\hat{\boldsymbol{\beta}}_Z).$$