Supplementary exercise 10.33 of IPS7e

Data & Model: For 40 children from Papua New Guinea, measurements of

- C-reactive protein (crp): indicative of infections,
- serum retinol: low values indicate vitamin A deficiency.

Both variables are response variables, and interest is in whether infections (indicated by crp) are associated with lower values of retinol. The question hypothesizes a causal effect, but a single observational study will never suffice to establish causality, so we will only discuss associations. The research question will lead to a linear regression model for serum retinol as a function of crp. The model assumes

$$\text{retinol}_i = \beta_0 + \beta_1 \text{crp}_i + \varepsilon_i, \;\; i = 1, \ldots, 40,$$

where the errors (ε_i) are i.i.d. from $\text{N}(0, \sigma)$.

(a) We use stemplots and graphical summaries from Minitab.

Stem-and-Leaf Display: retinol, crp

Stem-and-leaf of retinol N = 40

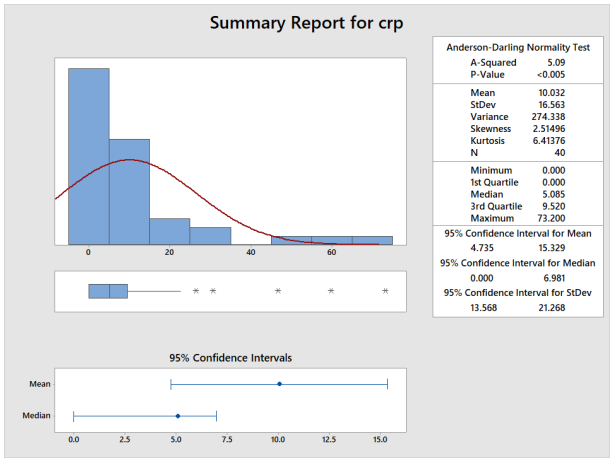
13	0	2333333333333
16	0	455
20	0	6667
20	0	88889999
12	1	00011111
4	1	23
2	1	4
1	1	
1	1	9

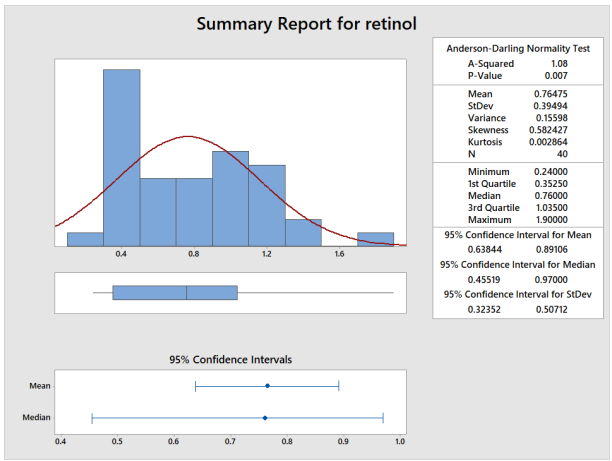
Leaf Unit = 0.1

Stem-and-leaf of crp N = 40

20	0	00000000000000003334
20	0	55555677899
9	1	2
8	1	5
7	2	02
5	2	6
4	3	0
3	3	
3	4	
3	4	6
2	5	
2	5	9
1	6	
1	6	
1	7	3

Leaf Unit = 1





Comments:
The distribution of crp is very strongly right-skewed, and 16 out of the 40 values equal 0. The distribution's right tail is quite long. The distribution of serum retinol also appears somewhat right-skewed, but it may also be bimodal with one mode around 0.35 and another mode around 0.9.

(b) No. There are no assumptions about the distribution of x in a linear regression. With a large number of x-values equal to zero there might be a concern about the equal variances or the linearity of the relation, but the fact that many x-values are equal to zero does not in itself violate the model's assumptions.

(c) We fit the regression in the Fitted line plot menu, and include residual plots.

Regression Analysis: retinol versus crp

The regression equation is
retinol = 0.8430 - 0.007800 crp

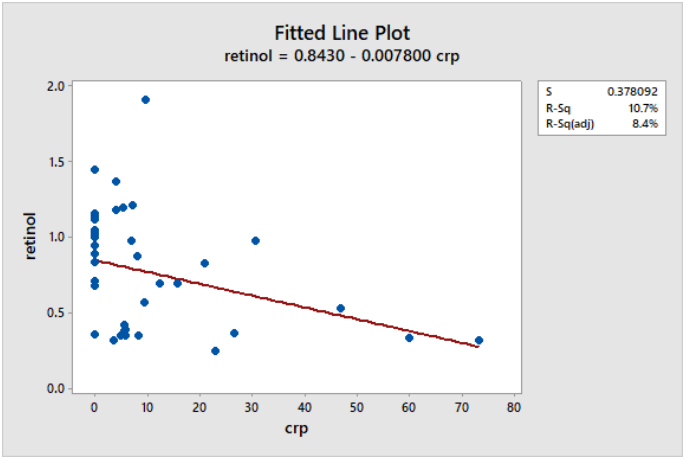
Model Summary

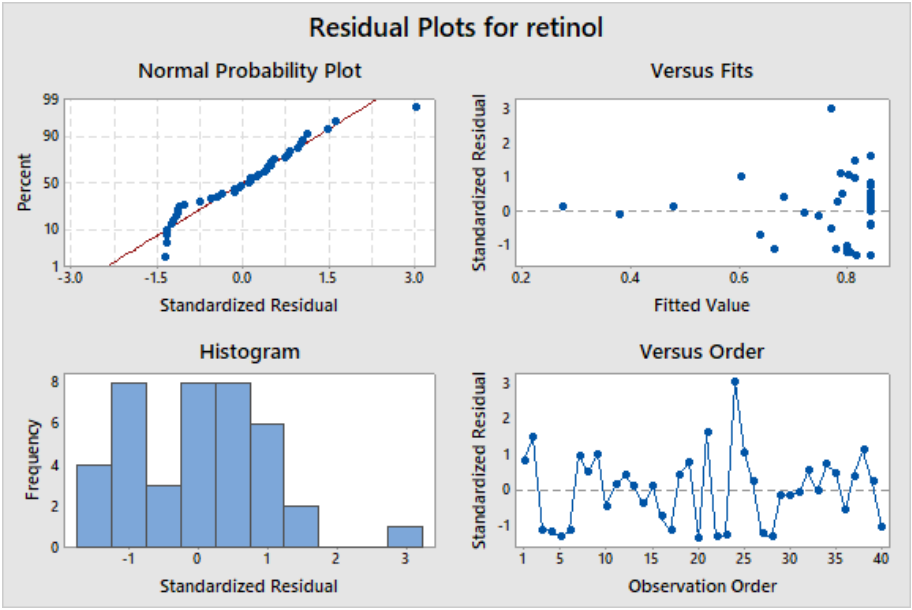
S	R-sq	R-sq(adj)
0.378092	10.70%	8.35%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.65097	0.650972	4.55	0.039
Error	38	5.43223	0.142953		
Total	39	6.08320			

Fitted Line: retinol versus crp





Comments:
The fitted line plot shows a very noisy relation. The fitted regression line is:

$$\text{retinol} = 0.8430 - 0.007800 \cdot \text{crp}.$$

Due to the large scatter of points around the line, it is hard to assess whether the association is linear. It is very weak, with $R^2 = 10.7\%$ and a barely significant test for $H_0 : \beta_1 = 0$ (i.e., a slope of zero). The estimated slope is negative, offering some support for the assumption that high crp-values could be associated with low serum retinol values. However, the weakness of the relation and the quite irregular scatter of points around the line suggest that caution must be exercised in interpreting these results.

(d) The assumptions of the linear regression model are (slide 11L−7):

- linear relation,
- normal distribution of errors,
- same standard deviation/variance for all errors (or observations),

– independence of the errors (or observations).

We already discussed the linear relation, and it is impossible to assess violations of the independence assumptions without further information about the data (which factors or circumstances that could possibly violate the independence).

We assess the assumed normal distribution for errors by residual plots. The normal plot for the standardized residuals looks strange, with a fairly narrow spread of the observations around zero except for one single large residual (obs. 24) around 3.0. The left tail of the distribution seems to be too short for a normal distribution. The fitted line plot gives no obvious explanation of this curious finding, but it can be seen that the points below the line are somewhat clustered at a moderate distance from the line instead of being nicely spread out at different distances from the line. As for the variance homogeneity, the residual plot (against fitted values) suggests that the largest variability could be associated with the largest fitted values, i.e. crp-values around 0. This impression might however occur because of the many crp-values at 0, and there are indeed only 3 larger values which are all fitted very closely by the line. These values are quite influential for the line, but at least their pattern is consistent.

In summary, despite the weak relation and strong scatter around the line there seems to be some systematic departures from the model assumptions. It is not clear how strongly this could affect conclusions. In any case, such a weak relation should not give rise to strong conclusions about the relation.