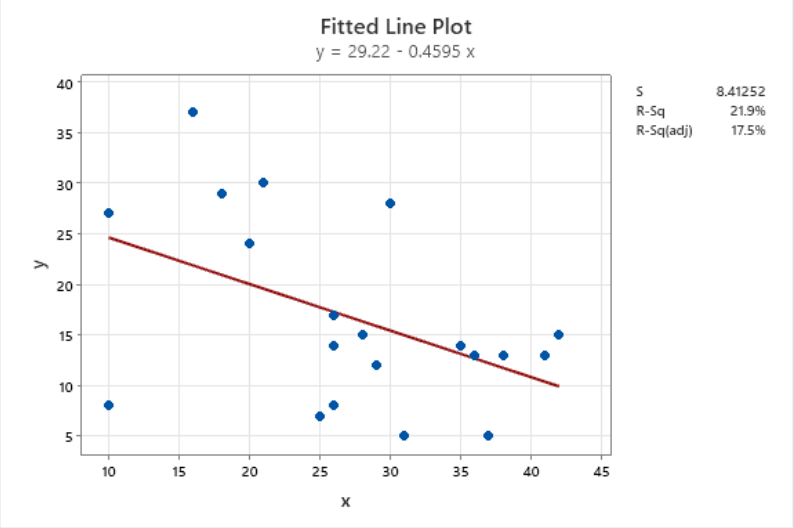
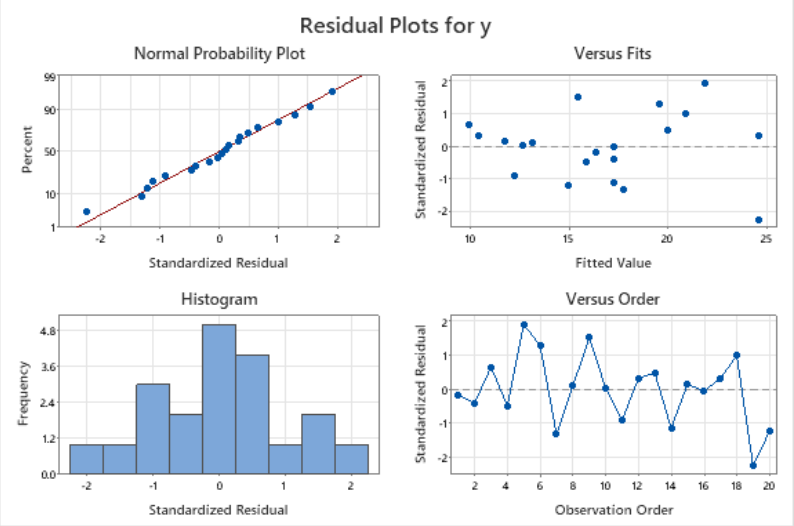
Additional Exercise 2.2

The analysis consists simply in fitting a linear regression model for a regression of y on x,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{where the } \varepsilon_i\text{'s are i.i.d. and} \sim \mathrm{N}(0, \sigma^2)$$

and to screen appropriate residuals and statistics for "strange things". The results shown below are from Minitab version 21, using the "Fitted Line Plot" menu to show the data points with the fitted line and to display the (standardised) residual plots in the "Four in one" layout.





Overall, the residual plots look nice and show nothing terribly suspicious. In order to get the table of regression coefficients and the table of "Unusual Observations", we repeat the analysis from the "Fit Regression Model" menu and also store the regression diagnostics available in the worksheet to be

displayed in a listing for the full dataset.

HS02_2.CSV

Regression Analysis: y versus x

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	29.22	5.89	4.96	0.000
x	-0.459	0.205	-2.24	0.038

Model Summary

S	R-sq	R-sq(adj)
8.41252	21.86%	17.52%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	356.3	356.33	5.04	0.038
Error	18	1273.9	70.77		
Total	19	1630.2			

Fits and Diagnostics for Unusual Observations

Obs	y	Fit	Resid	Std Resid	
19	8.00	24.63	-16.63	-2.25	R

R Large residual

Data

Row	x	y	FITS	SRES	TRES	HI	COOK	DFIT
1	28	15	16.3554	-0.16533	-0.16079	0.050333	0.000724	-0.03702
2	26	14	17.2744	-0.39953	-0.39001	0.050926	0.004283	-0.09034
3	42	15	9.9226	0.66607	0.65544	0.178907	0.048334	0.30595
4	29	12	15.8959	-0.47559	-0.46512	0.051815	0.006180	-0.10873
5	16	37	21.8692	1.92277	2.09624	0.124989	0.264049	0.79226
6	21	30	19.5718	1.28759	1.31325	0.073145	0.065418	0.36892
7	25	7	17.7338	-1.31116	-1.33980	0.053000	0.048106	-0.31696
8	35	14	13.1390	0.10703	0.10405	0.085587	0.000536	0.03183
9	30	28	15.4364	1.53586	1.60119	0.054481	0.067959	0.38435
10	36	13	12.6795	0.04006	0.03893	0.095364	0.000085	0.01264
11	37	5	12.2200	-0.90787	-0.90321	0.106325	0.049031	-0.31154
12	41	13	10.3820	0.33995	0.33144	0.162020	0.011172	0.14574
13	20	24	20.0313	0.49215	0.48154	0.081144	0.010695	0.14310
14	26	8	17.2744	-1.13164	-1.14110	0.050926	0.034358	-0.26433
15	38	13	11.7605	0.15693	0.15261	0.118471	0.001655	0.05595
16	26	17	17.2744	-0.03348	-0.03253	0.050926	0.000030	-0.00754
17	10	27	24.6262	0.32081	0.31266	0.226307	0.015052	0.16910
18	18	29	20.9503	1.00903	1.00957	0.100696	0.057001	0.33782
19	10	8	24.6262	-2.24689	-2.57422	0.226307	0.738352	-1.39223
20	31	5	14.9769	-1.22214	-1.24028	0.058332	0.046262	-0.30869

The fitted line plot shows the observation (10,8) in the lower left corner to be considerably off the regression line. It is exactly this observation (# 19) that shows in the Minitab table of Unusual Observations. The standardised residual is -2.25 (and the deletion residual is -2.57); neither of these are large enough to provide overwhelming evidence of data error. We can compute the outlier test based on the deletion residual as follows,

$$P(t(17) > -2.574) = 0.009855, \quad P\text{-value for outlier test } = 2 \cdot 20 \cdot 0.009855 = 0.39,$$

so there is no significance for it being an outlier. Nor is the leverage (0.23) very large (and it shouldn't be, because the x-value is not an outlier in the distribution of x's). The observation does show up as the clearly most extreme value on both Cook's D and DFITS (0.74 and -1.39, respectively), so there is some indication that it is influential. The reason why the statistics do not point more clearly to this point as an outlier is, that there is much noise about the regression line, with an R^2 of only 0.22 (increasing to 0.41 without the suspect observation, not shown).