Durham University
MATH1541 Statistics
Exercise Sheet 8

Kamil Hepak
Tutorial Group 4

Dec 2018

1 Find s_ϵ.

To find s_ϵ, we use the formula $s_\epsilon = s_y\sqrt{1-R^2}$. As per the R output, $R^2 = 0.925$ and $s_y = 4099.8$. Thus, $s_\epsilon = 1122.776$.

2 Assess the relative value of the predictors.

$Variable$	x_1	x_2	x_3	x_4	x_5	x_6
$\lvert\hat{b}_i\rvert s_i$	39.65	1363.50	5127.92	535.86	9636.55	14114.14

Using the relationship "value of variable $x_i \propto |\hat{b}_i|s_i$", we can see that Site 6 is the greatest contributor, and thus the most relevant predictor, by a fair margin. Sites 5 and 3 also contribute a large amount to the value of the prediction, Sites 2 and 4 contribute considerably less, and Site 1 contributes such an insignificant change that it may be a candidate for exclusion when computing $\hat{y}$. All 6 of the variables' standard deviations are sufficiently similar, thus we can fairly confidently compare them directly. Before excluding any variable, we would need access to data about the residuals for y - if their plots against any variable exhibited heteroscedasticity, then we may consider excluding that variable.

3 Predict the value of run-off volume, and find a 90% confidence interval.

To predict y, we use the formula:

$$\hat{y} = \hat{a} + \hat{b}_1 \cdot x_1 + \hat{b}_2 \cdot x_2 + \hat{b}_3 \cdot x_3 + \hat{b}_4 \cdot x_4 + \hat{b}_5 \cdot x_5 + \hat{b}_6 \cdot x_6$$

Following the R output and given values of x_{1-6}, the calculation is $\hat{y} = -12.8(7) - 664.4(4) + 2270.7(4) + 69.7(10) + 1916.5(10) + 2211.6(12) = 52736.8$. Assuming that, for these specific values of x_{1-6}, $y \sim N(52736.8, 1122.776^2)$, we can compute the 90% confidence interval for y as follows: $52736.8 \pm 1.6449(1122.776)$. This results in an interval of (50889.9, 54583.7).