ANOVA Formula

If Z and Y are jointly distributed (with finite second moments), then

$$E(Y) = E\left[E(Y|Z)\right],$$

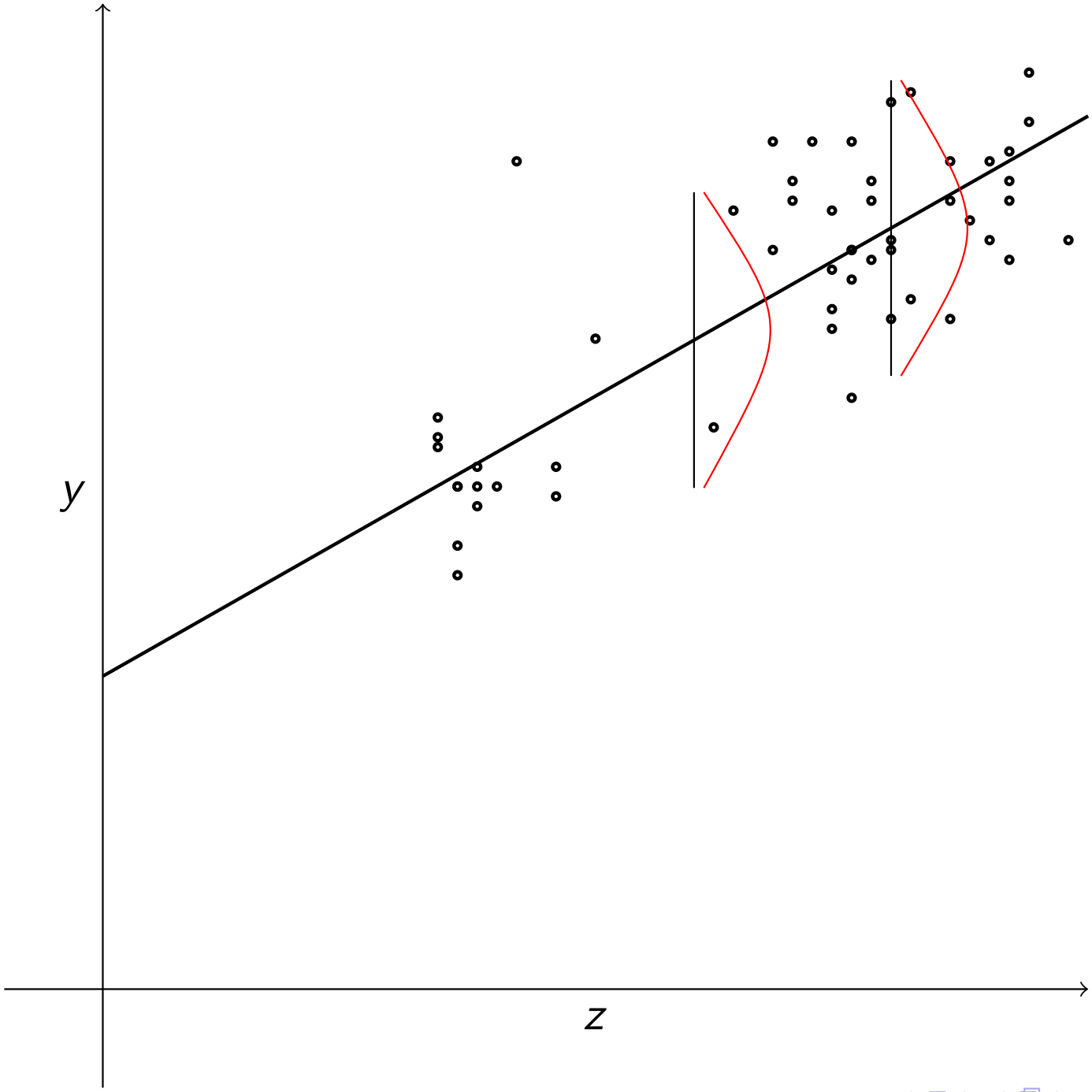
$$Var(Y) = E\left[Var(Y|Z)\right] + Var\left[E(Y|Z)\right] \geq Var\left[E(Y|Z)\right].$$

The first term on RHS is the 'within variation': if Y is partitioned according to values of Z, how much is left to be explained in Y for given Z. The second term is the variation between $\hat{Y}(Z)$ values, and is the 'between variation'. In a study, $Var(Y)$ may be large, but if $Var(Y|Z)$ is small, it makes sense to use Z to predict Y using Z. This result is known as the Analysis of Variance formula, and the ANOVA for regression is based on it.

z = duration and y = interval (both in minutes) for eruptions of Old Faithful Geyser

z	y	z	y	z	y	z	y	z	y	z	y
4.4	78	3.9	74	4.0	68	4.0	76	3.5	80	4.1	84
2.3	50	4.7	93	1.7	55	4.9	76	1.7	58	4.6	74
3.4	75	4.3	80	1.7	56	3.9	80	3.7	69	3.1	57
4.0	90	1.8	42	4.1	91	1.8	51	3.2	79	1.9	53
4.6	82	2.0	51	4.5	76	3.9	82	4.3	84	2.3	53
3.8	86	1.9	51	4.6	85	1.8	45	4.7	88	1.8	51
4.6	80	1.9	49	3.5	82	4.0	75	3.7	73	3.7	67
4.3	68	3.6	86	3.8	72	3.8	75	3.8	75	2.5	66
4.5	84	4.1	70	3.7	79	3.8	60	3.4	86		

Table: Eruptions of Old Faithful Geyser, August 1 – 4, 1978


y
z