EXERCISE 1

a) $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$

$i = 1, \dots, 13$ with $\varepsilon_i \sim N(0, \sigma^2)$ independent.

where x_{i2} = proportion of aluminum

x_{i3} = proportion of silicate

x_{i4} = proportion of aluminum-ferrite

x_{i5} = proportion of silicate-bic

b) ① "std error" of $\hat{\beta}_2$ (aluminium)

$t^{obs} = 3.435$ (from the table) is the observed value of the test

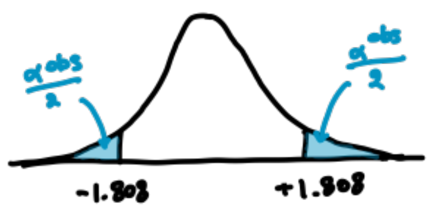
for testing $\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$

$$t^{obs} = \frac{\hat{\beta}_2 - 0}{\sqrt{\hat{var}(\hat{\beta}_2)}} = \frac{0.9739}{\text{std. error}} = 3.435 \Rightarrow \text{std. error}(\hat{\beta}_2) = \frac{0.9739}{3.435} = 0.2835$$

② t statistic of aluminum-ferrite

$$t^{obs} = \frac{\hat{\beta}_4 - 0}{\sqrt{\hat{var}(\hat{\beta}_4)}} = \frac{-0.4974}{0.2751} = -1.808$$

the pvalue "Pr(>|t|)" is



$$\alpha^{obs} = \mathbb{P}_{H_0}(|T| > |t^{obs}|)$$
$$= 2\,\mathbb{P}_{H_0}(T > |t^{obs}|) = 2\,\mathbb{P}_{H_0}(T > |-1.808|)$$

where $T \overset{H_0}{\sim} t_8$ (Student's t with $8 = n-p$ d.o.f.)

$$0.90 < \mathbb{P}(T \leq 1.808) < 0.95$$
$$-0.90 > -\mathbb{P}(T \leq 1.808) > -0.95$$
$$1-0.90 > 1-\mathbb{P}(T \leq 1.808) > 1-0.95$$
$$0.10 > \mathbb{P}(T > 1.808) > 0.05$$
$$0.20 > 2\,\mathbb{P}(T > 1.808) > 0.10 \Rightarrow \alpha^{obs} \in (0.10, 0.20)$$

③ estimate of the coeff. of silicate-bic

$$t^{obs} = \frac{\hat{\beta}_5}{\sqrt{\hat{var}(\hat{\beta}_5)}}$$

$$-2.481 = t^{obs} = \frac{\hat{\beta}_5}{0.3214} \Rightarrow \hat{\beta}_5 = -2.481 \cdot 0.3214 = -0.7974$$

④ $$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{13}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{13}(y_i - \bar{y})^2}$$

we have the error sum of squares: $SSE = 49.378$

and the total sum of squares: $SST = 2715.76$

the deviance decomposition is $SST = SSE + SSR$

$$\Rightarrow SSR = 2715.76 - 49.378 = 2666.385$$

Hence $R^2 = \frac{2666.385}{2715.76} = 0.9818$

If I consider, for example, a fixed significance level $\alpha = 0.05$, the statistically significant variables are the ones for which $\alpha^{obs} < 0.05$, hence, the intercept, aluminum, and silicate-bic.

c) $\begin{cases} H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1: \overline{H_0} \quad \text{(at least one is } \neq 0) \end{cases}$

under H_0 the model is $Y_i = \beta_1 + \varepsilon_i \rightarrow \tilde{\sigma}^2 = \frac{1}{13}\sum_{i=1}^{13}(y_i - \bar{y})^2 = \frac{1}{13} SST$

under H_1 the estimate of σ^2 is $\hat{\sigma}^2 = \frac{1}{13}\sum_{i=1}^{13}(y_i - \hat{y}_i)^2 = \frac{1}{13} SSE$

$$F = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \cdot \frac{n-p}{p-1} = \frac{SST - SSE}{SSE} \cdot \frac{13-5}{5-1} = \frac{SSR}{SSE} \cdot \frac{8}{4} \overset{H_0}{\sim} F_{4,8}$$

$$f^{obs} = \frac{2666.85}{49.378} \cdot 2 = 108.01$$

I reject H_0 if $f^{obs} > F_{4,8;1-\alpha}$

here, I reject at all significance levels α

d) model B: $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_5 x_{i5} + \varepsilon_i$

$SSE_B = 74.762$

$\begin{cases} H_0: \beta_3 = \beta_4 = 0 \\ H_1: \text{at least one of } (\beta_3, \beta_4) \neq 0 \end{cases}$

I use the statistic

$$F = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \cdot \frac{n-p}{p-p_0} \overset{H_0}{\sim} F_{p-p_0, n-p}$$

$\tilde{\sigma}^2$ is the estimate of σ^2 in model B $\quad \tilde{\sigma}^2 = \frac{1}{13} SSE_B$

$\hat{\sigma}^2$ is the estimate of σ^2 in model A $\quad \hat{\sigma}^2 = \frac{1}{13} SSE_A$

$p - p_0 = 5 - 3 = 2$

$n - p = 13 - 5 = 8$

$$\Rightarrow F = \frac{SSE_B - SSE_A}{SSE_A} \cdot \frac{8}{2} \overset{H_0}{\sim} F_{2,8}$$

$$f^{obs} = \frac{74.762 - 49.378}{49.378} \cdot 4 = 2.056$$

I reject H_0 if $f^{obs} > f_{2,8;1-\alpha}$ $\quad$ I do not reject H_0 for all usual α

I prefer model B since there is no evidence supporting the need to include x_3 and x_4, and a parsimonious model is preferable.

e) $$R^2_B = \frac{SSR_B}{SST} = \frac{SST - SSE_B}{SST} = 1 - \frac{74.762}{2715.763} = 0.9725$$

No, because the two models are nested and $R^2_A \geq R^2_B$ by construction.

f) residuals vs fitted : e_i vs $\hat{y}_i$

we use it to check if the model's assumptions are satisfied, specifically orthogonality of the residuals and homoscedasticity.

A "good" model would produce a plot without particular patterns

Here, the plot does not highlight troubling behaviors.

normal q-q plot

used to check the normality assumption: empirical vs theoretical quantiles

If the hyp. is satisfied, points should lie on the line

Here, we have few data to derive accurate conclusions.

However, the data do not show troubling patterns.