MVE 190 / MSG 500 Linear Statistical Models 2015-01-15

Question 1

a) Assume $\text{var}(\varepsilon_1) = \sigma_1^2$ and $\text{var}(\varepsilon_2) = \sigma_2^2$, independent, then

$$\begin{aligned}\text{var}(\hat{\beta}) &= \text{var}(w\hat{\beta}_1 + (1-w)\hat{\beta}_2) = w^2\,\text{var}(\hat{\beta}_1) + (1-w)^2\,\text{var}(\hat{\beta}_2) \\ &= w^2\sigma_1^2(X^TX)^{-1} + (1-w)^2\sigma_2^2(X^TX)^{-1} \\ &= (w^2\sigma_1^2 + \sigma_2^2 - 2w\sigma_2^2 + w^2\sigma_2^2)(X^TX)^{-1} \\ &= [(\sigma_1^2+\sigma_2^2)w^2 - 2w\sigma_2^2 + \sigma_2^2](X^TX)^{-1}\end{aligned}$$

$$\frac{\partial\,\text{var}(\hat{\beta})}{\partial w} = 2(\sigma_1^2+\sigma_2^2)w - 2\sigma_2^2 = 0 \iff$$

$$2(\sigma_1^2+\sigma_2^2)w = 2\sigma_2^2 \iff$$

$$w = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$

b) The experiment can help deciding which variables are important to include, but until the data is available and the diagnostics are run, we can't know if it will be necessary to reduce the variables.

Question 2

a) $\exp(-9.7939\text{42}) = 0.00005579$

b) When fixed at some particular value it can be used to compute the conditional logit of being in an honors class. It can also be interpreted as the difference in the log odds

c) Fix the math score at some value, say x, then

$-9.7939\text{42} + 0.1563\text{404}(x+1) - [-9.7939\text{42} + 0.1563\text{404}x]$

$= 0.1563\text{404}$

is the expected change in log odds.

Question 3

See lecture notes

Question 4

See lecture notes

Question 5

a) Top-right figure corresponds to CART. Top-left can't be generated with binary splits since each split of variable, say, X_1, partitions the entire space independently of X_2.

b) See lecture notes

Question 6

a) See lecture notes

b) AIC → M5

BIC → M2

C_p → M5

R^2_{adj} → M5

c) Cross-validation, backward or forward selection, lasso

d) It should be determined by the analysis goals. See lecture notes

Question 7

a) In principle, there's no problem with using this data (as long as one is confident in its quality). However, regression can't establish causality!

b) Rain acidity is a variable of a very different nature than the other variables included in the study. It should be independent of the others but spurious correlations can appear. Also, is it significant?

c) It depends on its interactions with other variables

d) Not particularly if it is correlated with production and there are no collinearity problems