

DSA1101: Introduction to Data Science

Practice Problems for 13 November, 2018

1. Suppose $\{(x_1, y_1, z_1), \dots, (x_n, y_n, z_n)\}$ are n observations on variables x , y and z , and we want to fit the regression model

$$y_i = \alpha + \beta x_i + \gamma z_i + \epsilon_i,$$

where ϵ_i is the error corresponding to the i^{th} observation.

- (i) Use the least squares method to deduce the equations which can be solved to obtain estimates of α , β and γ .
- (ii) Suppose $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ are the least squares estimates. Find the values of the following:

- (a) $\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i - \hat{\gamma}z_i)$
- (b) $\sum_{i=1}^n x_i (y_i - \hat{\alpha} - \hat{\beta}x_i - \hat{\gamma}z_i)$
- (c) $\sum_{i=1}^n z_i (y_i - \hat{\alpha} - \hat{\beta}x_i - \hat{\gamma}z_i)$

Use the decision tree to predict if each of the following five nursing homes is rural or



not.

BED	MCDAYS	TDAYS	PCREV	NSAL	FEXP	Rural (0/1)
82	114	136	8459	1995	1911	
123	208	336	22555	6059	6402	
120	238	363	17421	5362	6225	
78	154	203	9327	3672	1242	
58	166	205	10412	2245	1122	