

Exercises on Linear Models

1. Suppose $(x_i, y_i), i = 1, \dots, n$ is a set of pairs of observations, Consider the regressions of y on x and of x on y . Show:

- The regressions have slopes of the same sign.
- These give the same line if and only if $RSS = 0$ for both.
- The slope of the regression of x on y is in all other cases greater (in modulus) than the slope of the regression on y on x , and the ratio of the slopes is $0 \leq r^2 < 1$.
- Explain how (c) can be so, when in both cases $\hat{\beta}$ is an unbiased estimator of β .

2. The ‘hat’ matrix $H = X(X^T X)^{-1} X^T$ has a number of useful properties.

- Show that $H^2 = H$, $H(I - H) = 0$ and $(I - H)^2 = I - H$. (Both H and $I - H$ are said to be *idempotent*). Give a geometrical interpretation.
- Show that the eigenvalues of H are p ones and $n - p$ zeroes. [Hints: eigenvalues of H are also eigenvalues of H^2 ; you know the trace of H .]
- Show that $0 \leq h_{ii} \leq 1$.

3. Consider a transformation model of the form $\log(Y + \alpha) = x^T \beta + \epsilon$, $\epsilon \sim N(0, \sigma^2)$.

- Find an expression for the density of Y .
- Find an expression for the profile log-likelihood for the transformation parameter α given a sample y , that is

$$\hat{L}(\alpha; Y) = \sup_{\beta, \sigma^2 | \alpha} L(\alpha, \beta, \sigma^2; Y)$$

- How would you use this to choose a value of α ? Compare your suggestion with the example in Venables and Ripley (2002, p. 171).

4. Consider a multiple regression with $p \geq 2$ regressors. The F -test of the overall regression may be significant or not, and none, some or all of the t -tests for value zero of the coefficients b_i may be significant.

All six combinations of test results (F -test accept or reject, none/some/all t tests reject) can actually occur. The following example is designed to show this. Consider the model

$$y = \beta_1 + \beta_2(x + a) + \epsilon$$

where a is a constant and x is a vector of mean zero and variance one.

- (a) Calculate the F - and t -tests explicitly.
- (b) Show that by suitable choices of y and a we can construct datasets showing each of the six combinations.
- (c) Explain the practical implications of having a significant F -test but no significant t -test, and of all significant t -tests but no overall significance.

[For more on this see Largey & Spencer (1996) *The Statistician* **45**, 105–9.]

5. A set of data on 13 trials of cements is available:

x_1	x_2	x_3	x_4	y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

where y is the heat evolved in setting (in cal/gm) and $x_1 \dots x_4$ are the % of four components. The RSS of various models are:

model	RSS	model	RSS	model	RSS
----	2715.8	12--	57.9	123-	48.11
1---	1265.7	1-3-	1227.1	12-4	47.97
-2--	906.3	1--4	74.8	1-34	50.84
--3-	1939.4	-23-	415.4	-234	73.81
---4	883.9	-2-4	868.9		
		--34	175.7	1234	47.86

- (a) Choose a subset of regressors by forward selection.
- (b) Choose a subset of regressors by backward elimination.
- (c) Apply the stepwise selection procedure.
- (d) Compare the results of (a, b, c) with best subsets.
- (e) Explain why a small set of regressors suffices for this dataset.

[This is a famous set of data, originally from Hald, A. (1960) *Statistical Theory with Engineering Applications*. Wiley.]