Modelli Statistici Avanzato

Homework n. 5

With reference to Section 3.3-3.4 of the textbook of Fahrmeir, Kneib, Lang & Marx, briefly answer the following questions.

1. Suppose you fitted a model with an intercept and three covariates; considering the framework of general linear hypotheses, write the matrix ***C*** and the vector ***d*** to test the following hypotheses: (i) $5\beta_1 + 2 = 7\beta_2$; (ii) $\beta_1 = \beta_2 = \beta_3$
2. Suppose you have to test the hypothesis $\beta = 1$. Draw a plot like figure 3.15 on page 129 to show graphically (i) the *F* test; (ii) the Wald test (only the numerator). Briefly explain the difference between the principles underlying the *F* test and the Wald test.
3. Write the expression of the *confidence interval* for the mean of a new observation and the *prediction interval* for the value of a new observation; compute the difference between the length of the two intervals and comment.
4. Using matrix operations, derive the formula of the prediction interval in the case of a model with an intercept and a single regressor *x* (so that the design matrix *X* has two columns); note that the derived formula expresses the length of the prediction interval as a function of x_0, namely the value of the regressor for the new observation: which is the value of x_0 such that the prediction interval has minimum length?
5. Increasing the number of parameters of a model always increase the fit (reduce the sum of squared errors), nonetheless a model with more parameters is not necessarily preferable. Explain.
6. Suppose the true model is a linear model with design matrix **X**=[$\mathbf{X}_1$ $\mathbf{X}_2$]; however, you fit a model with only $\mathbf{X}_1$. Write the expectation of the estimator and discuss the bias: in which cases does the bias vanish?
7. Suppose you omit a relevant covariate, so that the estimator $\tilde{\beta}_j$ of the *j*-th regression coefficient is biased. Explain why in terms of $MSE(\tilde{\beta}_j)$ the biased estimator $\tilde{\beta}_j$ may be better than the unbiased estimator $\hat{\beta}_j$ based on the full model.
8. Write the definition of the sum of prediction squared errors (SPSE). Then write SPSE using the expression on the last line of page 145 and comment the meaning of the three components.
9. Suppose you take the sum of squared residuals as an estimator of SPSE: write the bias. Do you underestimate or overestimate SPSE? Is the bias more severe for a simple model or for a complex model?
10. Write the expressions of the AIC and BIC indexes and discuss similarities and differences. Then apply those indexes to choose between the following two models: model A has *M*=5 parameters and maximized log-likelihood *l*=–300, whereas model B has *M*=9 parameters and maximized log-likelihood *l*=–290. First suppose the dataset has size *n*=100. Then repeat assuming the dataset has size *n*=200. Comment.
11. Write the variance inflation factor (VIF) and explain why it is a useful diagnostic tool.
12. Write the ridge estimator and compare its properties with those of the least squares estimator.

Extra exercise to do with Stata

Replicate example 3.17 (page 141) of the textbook of Fahrmeir, Kneib, Lang & Marx. Note: use `set seed` to choose a value for the seed before randomly generating the values, then change the seed until you get results similar (not necessarily equal) to those in tables 3.3 and 3.4. Hint: to see how to randomly generate a dataset, look at the file `overfitting.do` in Moodle.