Imperial College London

Problem Sheet 9

MATH50011
Statistical Modelling 1

Week 10

Lecture 17: Inference with Normal Theory Assumptions

1. Consider the simple linear model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n.$$

Assume that the full rank and normal theory assumptions hold.

(a) Describe how to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at level α using (i) a t-test; and (ii) an F-test.

(b) Show that the p-values for the tests in (a) are equal.

(c) Derive a $(1 - \alpha) \times 100\%$ confidence interval for $E(Y|x_0)$, where x_0 is a fixed value of the covariate.

2. Suppose we believe that the distribution of Y depends on covariates x_1 and x_2, and that the relationship between Y and x_1 depends on the value of x_2. That is, we assume there is an *interaction* between x_1 and x_2.

To allow for the interaction, we use the following linear model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_1 x_2 + \epsilon_i.$$

The term $\beta_3 x_1 x_2$ is called an *interaction term*.

Now, assuming that the full rank and normal theory assumptions hold:

(a) Derive expressions for $E(Y_i|x_{i1} = x, x_{i2} = 0)$ and $E(Y_i|x_{i1} = x, x_{i2} = 1)$.

(b) State a hypothesis in terms of the parameter vector that could be used to test for the presence of an interaction between x_1 and x_2. Construct a level α test of your hypothesis, clearly identifying the form of the test statistic and its distribution under the null hypothesis.

(c) State a hypothesis in terms of the parameter vector that could be used to test for the presence of *any* effect of x_1 on the distribution of Y. Construct a level α test of your hypothesis, clearly identifying the form of the test statistic and its distribution under the null hypothesis.

Lecture 18: Outliers, Under- and Over-fitting, WLS

3. Let $Y = X\beta + \epsilon$ and assume that $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I$. Moreover, assume that X is an $n \times p$ matrix with full column rank. Let $\hat{\beta} = (X^T X)^{-1} X^T Y$ denote the least squares estimator of β under $E(Y) = X\beta$).

   (a) Suppose that you fit the model in which $E(Y) = X\beta$ when the true model is such that $E(Y) = X\beta + Z\gamma$. That is, the model is under fitted.

      i. Show that $\hat{\beta}$ is typically a biased estimator of β.

      ii. Under which conditions on Z we have that $\hat{\beta}$ is an unbiased estimator of β,

      iii. Compute $Cov(\hat{\beta})$.

   (b) Let $X = (X_1, X_2)$, where X_1 denotes the matrix with the first k columns of X. Suppose that you fit the model $E(Y) = X\beta$ when the true model is $E(Y) = X_1\beta$. That is, the model is over fitted.

      i. Show that the fitted model provides an unbiased estimator of the true model.

      ii. Show that the elements of $\hat{\beta}$ have in general higher variance than would result from fitting the true (reduced) model.

      iii. Under which conditions on X the elements of $\hat{\beta}$ do have higher variance than would result from fitting the true (reduced) model.

R lab: Hypothesis testing in linear models

4. The file psa.csv is a comma-separated file containing data on 28 men having hormonally treated prostate cancer. The first line of the file contains the following variable names. Each successive line contains data pertinent to one of the 28 patients.

nadirpsa = lowest PSA value attained post therapy (ng/ml)
grade = tumor grade (1= least aggressive, 3= most)
age = patient's age (years)
obstime = time in remission (months)

   (a) Define a linear model for $E(\log Y_i)$ where Y_i is the time spent in remission by the ith patient. In your model, include *nadirpsa*, *age*, as well as *grade* (treated as a 3-level categorical variable). How many parameters are in your model?

   (b) Using the software of your choice, obtain the least squares estimates corresponding to the linear model you specified in part (a).

   (c) Is there evidence that *nadirpsa* is associated with time in remission? Implement a hypothesis test to justify your conclusion.

   (d) Modify the linear model in (a) to allow for an interaction between tumor grade and age. How many parameters are in your model?

   (e) Using the software of your choice, obtain the least squares estimates corresponding to the linear model you specified in part (d).

   (f) Is there evidence of an interaction between tumor grade and age? Implement a hypothesis test to justify your conclusion.