LINEAR MODELS; THE SIMPLE LINEAR REGRESSION MODEL

1. Linear models

Suppose we have a set X, which may be the set of real numbers, and some real-valued functions $f_0 \equiv 1$ and $f_1, ..., f_k$ which are linearly independent, meaning that there are no constants $c_0, c_1, .., c_k$, not all 0, such that $\sum_{i=0}^{k} c_i f_i \equiv 0$. Let $x_1, ..., x_n$ be some points of X, with $k < n$. We observe some random variables $Y_1, Y_2, ..., Y_n$. The *linear model* for (x_j, Y_j), $j = 1, ..., n$, based on $f_1, ..., f_k$ is that for some $\varepsilon_1, ..., \varepsilon_n$ i.i.d. $N(0, \sigma^2)$ for some unknown $\sigma > 0$, for some unknown real numbers $a_1, ..., a_k$,

$$Y_j = \sum_{i=0}^{k} a_i f_i(x_j) + \varepsilon_j. \tag{1}$$

"Linear model" means that the model is linear with respect to the coefficients $a_0, a_1, ..., a_k$. If X is a vector space, such as the real line, the functions f_i need not be linear on it.

In R, one can fit the linear model (1) to data (x_j, Y_j) for $j = 1, ..., n$, where x_j are fixed non-random design points, by a command of the form

$$\text{regrobj} = \text{lm}(Y \sim f_1, ..., f_k) \tag{2}$$

where "regrobj" can be replaced by any name, not already defined in the R system, one chooses to give the regression "object". Then summary(regrobj) will output estimates $\hat{a}_i$ of the coefficients a_0 ("Intercept") and a_i (coefficient of f_i) for $i = 1, ..., k$. Here $\widehat{a}_i$ are random variables under the model, as they depend on the random ε_j. These random variables will be written $\widehat{a}_i^{rv}$. On the other hand for given observed $Y_1, ..., Y_n$, solving for the estimates $\widehat{a}_i$ gives constants that will be called $\widehat{a}_i^{obs}$. The *p-value* of the estimate $\widehat{a}_i$ is the probability under the model with $a_i = 0$, of obtaining as large a value $\widehat{a}_i$ as the one observed,

$$\Pr(|\widehat{a}_i^{rv}| \geq |\widehat{a}_i^{obs}|).$$

If the p-value of $\widehat{a}$ is less then 0.05, one rejects the hypothesis that $a_i = 0$ and decides that $a_i \neq 0$.

Date: 18.650, Sept. 23, 2015.

In the multiple regression, R outputs p-values for the hypothesis that $a_i = 0$ for each i. R computes the p-values based on the assumption that the errors ε_j are i.i.d. $N(0, \sigma^2)$, which leads to t distributions.

2. Simple and quadratic regression models

In the *simple linear regression model*, X is the real line, $k = 1$, and $f_1(x) \equiv x$. Thus a straight line $y = a + bx$ or equivalently $a_0 + a_1 x$ is fitted to data (x_j, Y_j), $j = 1, ..., n$, in fact performing y-on-x regression.

In the *quadratic regression model* we have the larger model with $k = 2$ and $f_2(x) = x^2$. If the simple linear regression model is valid, then so is the quadratic regression model, with $a_2 = 0$. If we find in quadratic regression that the hypothesis $a_2 = 0$ has a small p-value (say, less than 0.05) then we can reject the simple linear regression model. In R, the special case of (2) for quadratic regression is written for example as

qobj = lm(Y∼ x+I(x^2))

where $I(\cdot)$ is the identity function but R requires it to be written.

3. Residuals

Once one has fitted a linear model 1, then one has estimates $\widehat{a}_i$ of the coefficients a_i for $i = 0, 1, ..., k$. The *residuals* are the quantities $\hat{\varepsilon}_j = Y_j - \sum_{i=0}^{k} \widehat{a}_i f_i(x_j)$. The "summary" gives a few of the residuals: the smallest, the lower quartile, the median, the upper quartile, and the largest (we haven't defined quartiles or even median yet in the course). By the command

residuals(regrobj)

one gets a list of the residuals $\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n$. If the x_i satisfy $x_1 < x_2 < \cdots < x_n$ as they normally should, then if the residuals appear "random" it seems that the regression has worked well. If a pattern, such as a convex (like x^2) or concave (like $-x^2$) one, is visible in the residuals, then the regression does not fit the data well. If the residuals are for a simple linear regression, it seems a quadratic regression might fit the data better.