Multiple Regression

1. Hypertension: How the relationship between the blood-pressure levels of newborns and blood-pressure levels of infants relates to subsequent adult blood pressure.
2. The blood pressure of newborn is affected by several extraneous factors that make this relationship difficult to study. In particular, newborn blood pressures are affected by birthweight the day of life on which blood pressure is measured.
3. In this study, the infants were weighed as the time of the blood-pressure measurements. This birthweight is different from actual birthweight. Birthweights at 5 days are different from those at 2 days. We want to adjust the observed blood pressure for these two factors before we look at other factors.

Model

$$y_i = \alpha + \sum_{i=1}^{k} \beta_j x_{ij} + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2)$$

F test for simple linear regression

To test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs. H_1 : at least one $\beta_j \neq 0$. In general consider the problem where two models, 1 and 2, where model 1 is 'nested' within model 2. Model 1 is the Restricted model, and Model 2 is the Unrestricted one. That is, model 1 has p_1 parameters, and model 2 has p_2 parameters, where $p_2 > p_1$. The model with more parameters will always be able to fit the data at least as well as the model with fewer parameters. Thus typically model 2 will give a better (i.e. lower error) fit to the data than model 1. But one often wants to determine whether model 2 gives a significantly better fit to the data. One approach to this problem is to use an F test.

If there are n data points to estimate parameters of both models from, then one can calculate the F statistic, given by

$$F = \frac{(RSS_1 - RSS_2)/(p_2 - p_1)}{RSS_2/(n - p_2)}$$

where RSS_i is the residual sum of squares of model i. Under the null hypothesis that model 2 does not provide a significantly better fit than model 1, F will have an F distribution,

with $(p_2 - p_1, n - p_2)$ degrees of freedom. The null hypothesis is rejected if the F calculated from the data is greater than the critical value of the F-distribution for some desired false-rejection probability (e.g. 0.05). The F-test is a Wald test.

T Test for an individual independent variable

$H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$.
Compute $t = \hat{\beta}_j / SE(\hat{\beta}_j)$ which follows a t distribution with $(n - k - 1)$ degrees of freedom under H_0. If $t > t_{n-k-1,1-\alpha/2}$ or $t < -t_{n-k-1,1-\alpha/2}$, then reject H_0, otherwise accept H_0.
The exact p-value is given by

$$\begin{cases} 2P(t_{n-k-1} > t) \,\text{if}\, t > 0 \\ 2P(t_{n-k-1} < t) \,\text{if}\, t < 0 \end{cases}$$

Partial residual plot

A partial-residual plot characterizing the relationship between the dependent variable y and a specific independent variable x_i in a multiple-regression setting is constructed as follows:

1. A multiple regression is performed of y on all predictors other than x_i (i.e., $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$), and the residuals are saved.
2. A multiple regression is performed of x_i on all other predictors (i.e., $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$) and the residuals are saved.
3. The partial-residual plot is a scatter plot of the residuals in step 1 on the y-axis against the residual in step 2 on the x-axis.

The partial-residual plot reflects the relationship between y and x_i after each variable is adjusted for all other predictors in the multiple-regression analysis.