1 Testing hypotheses in regression

There are three general types of hypotheses that we may be interested in testing in multiple regression: 1) $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$, 'None of the x variables explain variation in y' (Overall regression F test), 2) $H_0 : \beta_j = 0$, 'x_j does not explain additional variation in y beyond what is already explained by $x_1, x_2, .., x_{j-1}, x_{j+1}, ., x_k$' (Partial F test = standard t test), or 3) $H_0 : \beta_{g+1} = \beta_{g+2} = ... = \beta_k = 0$, ' $x_{g+1}, x_{g+2}, ..., x_k$ do not explain additional variation beyond what is already explained by $x_1, x_2, ..., x_g$' (Multiple Partial F test). All three of these kinds of hypotheses may be tested using the extra sum of squares principle. Using the extra sum of squares principle, we specify a full model, which includes all terms; and a reduced model, which consists of only those terms that remain if H_0 is true. These models are nested, and thus the full model will always have a larger SS(Regression) than the reduced model. The key question is if the SS(Regression) is increased enough in the full model to justify the use of the extra terms in the model. The null hypothesis H_0 can then be tested using the test statistic:

$$F = \frac{(SS(\text{Regression})_F - SS(\text{Regression})_R)/(\#\text{terms in } H_0)}{MS(\text{Residual})_F},$$

where the subscript R indicates the reduced model and the subscript F indicates the full model. The statistic F can be compared to the critical value $F_{df1,df2,\alpha}$ where $df1 = \#$terms in $H_0, df2 =$ the degrees of freedom for $MS(\text{Residual})_F$, and α is the prechosen significance level.

For example, with our sandwich data (excluding the two burgers) we have $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where $y =$ calories, $x_1 =$ fat, and $x_2 =$ cholesterol. We can test null hypothesis 1, $H_0 : \beta_1 = \beta_2 = 0$, by defining the full model to be $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ and the reduced model to be $y = \beta_0 + \varepsilon$. Check the SAS program to see how to specify the reduced model. Our results are: $SS(\text{Regression})_F = 6989.48$, $SS(\text{Regression})_R = 0.$, and $MS(\text{Residual})_F = 116.56$. Then our F statistic is

$$F = \frac{(6989.48 - 0.)/2}{116.56} = 29.98,$$

which can be compared to an $F_{2,4,\alpha}$ value. Note that this is the overall F test presented in the SAS printout, and has a P-value of .0039. It also turns out that the F test for hypothesis 2) $H_0 : \beta_j = 0$, is equivalent to the t-test next to the coefficient on the SAS printout. For hypotheses of the form 3) (Multiple Partial F test), we need to run the full and reduced models separately and use both outputs to conduct the test.

2 Regression diagnostics: residual analysis

Recall that the residuals ε in the multiple regression model $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon$ should 1) have a mean of 0, 2) be independent, 3) have a normal distribution, and 4) have a common variance σ^2. We can estimate the ε terms from our sample via one of several types of residuals:

Name of residual	**Formula**
residual	$\widehat{\varepsilon}_i = y_i - \widehat{y}_i$
standardized residual	$z_i = \widehat{\varepsilon}_i / s_\varepsilon$
studentized residual	$r_i = \widehat{\varepsilon}_i / (s_\varepsilon \sqrt{1 - h_i})$
jackknife residual	$r_{(-i)} = \widehat{\varepsilon}_i / (s_{\varepsilon(-i)} \sqrt{1 - h_i})$

where $s_\varepsilon = \sqrt{MSE}$, h_i is called the leverage of the ith observation and satisfies $0 \leq h_i \leq 1$, and $s_{\varepsilon(-i)}$ is s_ε computed without the ith observation. The last three types (standardized, studentized, and jackknife) are fairly similar if the regression assumptions are satisfied. Since the last type, the jackknife residual, is best at detecting various problems, we will often use it. It can be obtained in SAS Proc Reg on the Output statement with the RSTUDENT option. A sound strategy for diagnosing the adequacy of a regression model is to try a candidate model, then obtain a variety of plots of residuals. These include histograms, normal plots, and scatter plots of the residuals versus the predicted values, and the individual x_i values. Plots of residuals versus time can be useful for data that are collected over time, and an improved type of plot of residuals versus x_i, called a partial regression plot, adjusts for all other $x_{i'}$ terms in the model. Partial regression plots can be obtained in SAS Proc Reg by using the / PARTIAL option. A useful way to evaluate the collection of residual plots is to look at the plots for problems, and focus on characterizing the potential effect of these problems on the final regression model.