Multiple Linear Regressions

We often have more than one predictor. The reason we don't separate a model into n simple linear regressions because it is unclear how to make a single prediction given all the responses and variables. Also, each of the regression equations ignores the other (n-1) variables in forming estimates for the regression coefficients. This can lead to very misleading estimates of the association between variables and the responses. Instead of fitting separate linear regression model for each predictor, a better approach is to extend the simple linear equation regression model so that it can directly accommodate multiple predictors. The multiple linear regression model takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response. β_j is interpreted as the *average effect* on Y of a one unit increase in X_j, holding all other predictors fixed.

Estimating the Regression Coefficients

estimates:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

We choose $\beta_0, \beta_1, \ldots, \beta_p$

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The values that **minimize the RSS** are the multiple least squares regression coefficent estimates

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. Do all predictors help to explain Y, or is only a subset of the predictors useful.
3. How well does the model fit the data?
4. How accurate is our prediction??

1. Relaionship between the Response and Predictors? **Testing the null hypothesis**

   $$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

   versus the alternative

   $$H_a : \text{at least one } \beta_j \text{ is non-zero}$$

   This hypothesis test is performed by computing the *F-statistic*

   $$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}$$

   $\text{TSS} = \sum (y_i - \overline{y})^2$ and $\text{RSS} = \sum (y_i - \hat{y}_i)^2$

   $$E\{\text{RSS}/(n-p-1)\} = \sigma^2$$

   and if H_0 is true,

   $$E\{(\text{TSS} - \text{RSS})/p\} = \sigma^2$$

   F-statistic is expected to be close to 1.
   if H_a is true, then $E\{(\text{TSS} - \text{RSS})/p\} > \sigma^2$ and hence $F \gg 1$. Large F-statistic suggests that at least one of predictors are related to the responses.

For a large n, $F-$statistic > 1 might provide evidence against H_0 A larger F-statistic is needed to **reject** H_0 for a small n.
Null hypothesis test for a particular subset of q of the coefficients are zero.

$$H_0 : \beta_{p-q-1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$

Next: fit a second modelthat uses all the variables *except* those last q.

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)}$$

In this case, t-statistic and a p-value are both equivalent to the one from last model hence indicating *partial effect* of adding that variable to the model

NOTE: Using individual t-statistics and associated p-values in order to decide whether or no there is any association between the variables and the response, there is avery high chance that we will incorrectly conclude that there is a relationship. F-statistic does not suffer from this problem because it adjusts for the number of predictors. Hence, if H_0 is true, there is only a 5% chance that the F-statistic will result in a p-value below 0.05, regardless of the number of predictors or the number of observations. When p is large, **forward selection** is an approach used.

2. Deciding on **Important Variables**, after computing the F-statistic
Variable Selection: The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors

    1. **Forward Selection**. [Greedy Approach and might include redudant variables early on]
    • start with a **Null Model** (contains an intercept but no predictors)
    • fit p simple linear regressions
    • add to the null model the variable that results in the lowest RSS for the new two-variable model
    • continue till some stopping rule is satisfied

    2. **Backward Selection**. [Cannot be used for $p > n$]
    • start with all variables in the model
    • remove the variable with the largest p-value (variable that is the least statistically significat)
    • new (p-1) model, it's fit
    • again the variable with the largest p-value is removed.
    • continue till some stopping rule is satisfied (some threshold p-value)

    3. **Mixed Selection**. [Combination of forward and backward selection.]
    • start with no variables in the model
    • [forward selection] add the variable that provides the best fit
    • continue adding variables one-by-one
    • if at any point the p-value $>$ threshold for one of the variables
    $\rightarrow$ remove that variable from the model
    • continue to peform these forward and backward steps until all variables in the model have a sufficiently low p-value and all the variables outside the model would have a large p-value if added to the model.

3. **Model Fit** RSE and R^2

$$\text{RSE} = \sqrt{\frac{1}{(n-p-1)}\text{RSS}}$$

synergy or *interaction* effect between predictors and responses.

4. **Predictions** after fitting the multiple regression model, it is straightforward to apply in order to predict the response Y on the basis of a set of values for the predictors $X_1, X_2, \ldots, X_p$. However, there are three sorts of uncertainty associated with this prediction.

   1. The coefficients estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are estimates for $\beta_0, \beta_1, \ldots, \beta_p$ *Least squares plane*

      $$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

      is an estimate for the *true population regression plane*

      $$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

      We can compute a confidence interval in order to determine how close $\hat{Y}$ will be to $f(X)$

   2. **Model Bias** there is an additional source of potentially reducible error. When we use a linear model, we are in fact estimating the best linear approximation to the total surface. This discrepancy is ignored and we pretend that the linear model was correct ⌣

   3. The random error ε makes it impossible to predict the response value perfectly since it is an *irreducible error*. Prediction intervals- wider than confidence intervals because they incorporate both the error in the estimate for $f(X)$ (The reducible error) and the uncertainty as to how how much an individual point will differ from the population regression plane (the reducible error).