## Choosing the best model for prediction

Calculating the PRESS residuals $\delta_i^* = y_i - y_{i,-i}$ is an example of doing leave one out cross validation.
These residuals measure how well the i-th observation of the response can be estimated or predicted in terms of the other observations. The model with the best prediction ability should be the one that minimizes $\sum_{i=1}^{n} |\delta_i^*|$ or $\sum_{i=1}^{n} (\delta_i^*)^2 = \text{PRESS}$, ~~to~~ eventually the one that maximizes

$$R^2_{pred} = 1 - \frac{\sum_{i=1}^{n} (\delta_i^*)^2}{SS_T}$$

## Categorical or Indicator variables

Suppose $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 z_i + \varepsilon_i$
where $z$ is a categorical variable with $l$ categories.
Such regression problems are normally best solved by introducing $l-1$ indicator variables.
For instance, with three different catalyst the regression model can be:

$$\cancel{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 z_{i1} + \beta_4 z_{i2} + \varepsilon_i, \quad i = 1, 2, \ldots, m}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 z_{i1} + \beta_4 z_{i2} + \varepsilon_i, \quad i = 1, 2, \ldots, m$$

Where the colums $\underline{z}_1$ and $\underline{z}_2$ are

$$
\begin{array}{cc}
z_1 & z_2 \\
\left[\begin{array}{c|c}
1 & 0 \\
1 & 0 \\
\vdots & \vdots \\
1 & 0 \\
0 & 1 \\
0 & 1 \\
\vdots & \vdots \\
0 & 1 \\
0 & 0 \\
\vdots & \vdots \\
0 & 0
\end{array}\right]
\end{array}
$$

The $(1,0)$ combination corresponds to analyst 1, the $(0,1)$ to Catalyst 2 and the $(0,0)$ to Catalyst 3.

The model is then.

$$Y_i = (\beta_0 + \beta_3) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad \text{for catalyst } 1$$

$$Y_i = (\beta_0 + \beta_4) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad \text{—"— } 2$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \quad \text{—"— } 3$$

It may happen that the coefficients for the continuous variables depend on the ~~level of the~~ category variables.

Suppose we construct the model

$$\cancel{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 Z_{i1} + \beta_4 Z_{i2} + \beta_5 X_{i1} Z_{i1}}$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 Z_{i1} + \beta_4 Z_{i2} + \beta_5 X_{i1} Z_{i1} + \beta_6 X_{i2} Z_{i2} + \varepsilon_i$$

$i = 1, 2, \ldots, n$

Then we have:

$$Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad \text{catalyst } 1$$

$$Y_i = (\beta_0 + \beta_4) + \beta_1 X_{i1} + (\beta_2 + \beta_6) X_{i2} + \varepsilon_i \quad \text{catalyst } 2$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad \text{catalyst } 3.$$

# Orthogonal columns in the Design matrix

Let the design matrix $\underline{X} = [\underline{1}, \underline{x}_1, \underline{x}_2, \dots, \underline{x}_k]$

If $\underline{x}_p^T \underline{x}_q = 0$ i.e $\sum_{i=1}^{n} x_{ip} x_{iq} = 0$, $p \neq q$ $\underline{x}_p$ and $\underline{x}_q$ are orthogonal

If all columns are orthogonal $\underline{X}^T \underline{X} = \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & \underline{x}_1^T \underline{x}_1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & \underline{x}_k^T \underline{x}_k \end{bmatrix}$.

and $\hat{\underline{\beta}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{y} = \begin{bmatrix} 1/n & 0 & \cdots & 0 \\ 0 & (\underline{x}_1^T \underline{x}_1)^{-1} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & (\underline{x}_k^T \underline{x}_k)^{-1} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \underline{x}_1^T \underline{y} \\ \vdots \\ \underline{x}_k^T \underline{y} \end{bmatrix}$

The estimator for $\beta_1$ will only depend on $\underline{x}_1$ and $\underline{y}$

and $SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (b_0 + b_1 x_{i1} + \dots + b_k x_{ik} - b_0)^2$

$= b_1^2 \sum_{i=1}^{n} x_{i1}^2 + b_2^2 \sum_{i=1}^{n} x_{i2}^2 + \dots + b_k^2 \sum_{i=1}^{n} x_{ik}^2$

$= SS_R(\beta_1 | \beta_0) + SS_R(\beta_2 | \beta_0) + \dots + SS_R(\beta_k | \beta_0)$

## Testing a general linear hypothesis

Suppose $E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and we want to test,

$H_0: \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_2 - \beta_3 \end{bmatrix} \begin{matrix} = 0 \\ = 0 \end{matrix}$ or $H_0: \beta_1 = \beta_2 = \beta_3 = \beta$.

Can be formulated as: $H_0: \underline{C}\underline{\beta} = \underline{0}$ $\quad H_1: \underline{C}\underline{\beta} \neq \underline{0}$

where $C = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$ and $\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$, $\underline{C}\underline{\beta} = \underline{0} \Leftrightarrow \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_2 - \beta_3 \end{bmatrix} = \underline{0}$

We have $r$ (in this case 2) constraints on $\underline{\beta}$.

$$\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1}) \Rightarrow C\hat{\beta} \sim N(C\beta, \sigma^2 C(X^TX)^{-1}C^T)$$

Hence under $H_0$ $(C\hat{\beta})^T(\sigma^2 C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta}) \sim \chi^2_r$

Since $\hat{\beta}$ and $SS_E$ (full model) are independent and $\frac{SS_E}{\sigma^2} \sim \chi^2(n-k-1)$

we get.

$$F = \frac{(C\hat{\beta})^T(C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta})/\sigma^2 r}{SSE/\sigma^2(n-k-1)} \sim F_{r,\, n-k-1}$$

Under $H_0$. Reject if, $F$ is large.

Testing $H_0: C\beta = d$ $\quad H_1: C\beta \neq d$

we just substitute $C\hat{\beta}$ with $C\hat{\beta} - d$ in $F$