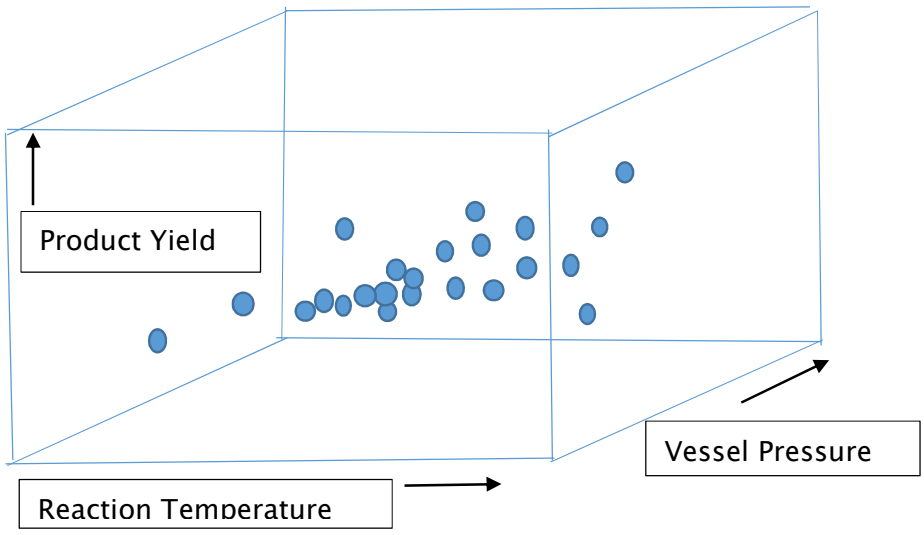
# A discussion on multiple regression models

In our previous discussion of simple linear regression, we focused on a model in which one independent or explanatory variable $X$ was used to predict the value of a dependent or response variable $Y$. We find such model is very useful in establishing a calibration curve for analytical instrument where the $X$ variable is the standard solution concentration and $Y$, the instrument response. However, in many other scientific studies, we encounter several independent variables which affect the final outcome of the experiment. Hence, we now want to model the dependent variable by several independent variables.

For example, we may want to carry out a drug chemical synthesis experiment by predicting its yield (g) based on its reaction temperature (°C) and vessel pressure exerted (psi). In this case, we have two independent variables to predict the value of the dependent variable, i.e. the drug yield. With these two dependent variables in the multiple regression model, a scatter diagram of the drug yield as an outcome variable can be plotted on a three-dimensional graph such as the diagram below:




In general, if $Y$ is the dependent or outcome variable and $X_1, X_2, \ldots, X_k$ are $k$ independent variables, then the general multiple regression model has the general form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

where

$\beta_o$ is $y$-intercept, and

$\beta_k$ is the slop of $y$ with variable $x_k$ when the other $X$'s are held constant.

The part $E(Y) = \beta o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ is the deterministic portion of the model. The $\varepsilon$ term is the random error in $Y$

Before further discussion, we have to take note of the assumptions made on multiple regression. The assumptions of simple regression discussed earlier also hold true for multiple regression. They are mainly:

1. For any given set of values of the independent variables, the random error $\varepsilon$ has a normal distribution with mean = 0 and standard deviation equal to $\sigma$;
2. The random errors are independent.

Furthermore, as soon as we use more than one independent variable, we have to worry about *multi-collinearity*. This means that none of our independent variables (also called predictors) should correlate highly with any other independent variable. In particular, no variable can be a linear combination of other variables, i.e. we cannot include as predictors the independent variables $A$, $B$ and $A+B$. Predictor variables that are highly correlated tend to explain much the same as variance in an outcome variable, blurring the relationship of each individual predictor with the outcome.

Let's see how the first-order model works in estimating and interpreting the parameters involved.

As in the case of simple linear regression, we also adopt the sample regression coefficients ($b_o$, $b_1$ and $b_2$) used as estimates of the population parameters ($\beta_o$, $\beta_1$ and $\beta_2$). Therefore, the regression equation for a multiple linear regression model with two explanatory variables, for example, is expressed as follows:

$$y = b_o + b_1 x_1 + b_2 x_2$$

For illustration, a series of experiments was carried out to synthesize a drug chemical by studying its total yield against two independent variables, reaction temperature and vessel pressure. The test results were summarized as below:

| Pressure, *psi* | Temperature, $^oC$ | Yield, *g* |
|---|---|---|
| 10 | 40 | 25 |
| 20 | 40 | 30 |
| 30 | 40 | 33 |
| 10 | 60 | 32 |
| 20 | 60 | 38 |
| 30 | 60 | 40 |
| 10 | 80 | 45 |
| 20 | 80 | 47 |
| 30 | 80 | 48 |

By applying the Excel software through **Data ->Data Analysis -> Regression**, we get the following Excel output with the values of the coefficients shown:

**SUMMARY OUTPUT**

| Regression Statistics | |
|---|---|
| Multiple R | 0.985 |
| R Square | 0.971 |
| Adjusted R Square | 0.961 |
| Standard Error | 1.602 |
| Observations | 9 |

ANOVA

| | df | SS | MS | *F* | Significance *F* |
|---|---|---|---|---|---|
| Regression | 2 | 510.833 | 255.417 | 99.585 | 2.501E-05 |
| Residual | 6 | 15.389 | 2.565 | | |
| Total | 8 | 526.222 | | | |

| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | 5.222 | 2.417 | 2.161 | 0.074 | -0.692 | 11.137 |
| Pressure | 0.317 | 0.065 | 4.843 | 0.003 | 0.157 | 0.477 |
| Temperature | 0.433 | 0.033 | 13.256 | 0.000 | 0.353 | 0.513 |

The estimated regression equation therefore is

$$Yield = 5.22 + 0.37\ Pressure + 0.433\ Temperature$$

The above equation tells us that adding one psi vessel pressure adds 0.37g drug yield and increasing one degree reaction temperature increases the drug yield by 0.43 g. In other words, if the reaction temperature was fixed at 80°C and the vessel pressure, at 20 psi, the drug chemical yield might be:

$$Yield = 5.22 + 0.37\,(20) + 0.433\,(80) = 46.2\ g$$

We can make few inferences from the parameters obtained, such as:

1. Estimating the confidence interval on $\beta_1$, the coefficient in front of the pressure variable in the above example.

In general, the confidence interval for βi is $b_i \pm t_{\alpha/2}$ x *standard error of* $b_i$. The degrees of freedom for the *t*–value is $n-(k+1)$, where $n$ = sample size, $(k+1)$ = the number of *beta* in the model.

Hence, in this case, we have *9 – (2+1) = 6* degrees of freedom and the critical $t_{0.5/2} = 2.447$. The 95% confidence interval for the pressure coefficient, $\beta_1$ is given by $b_1 \pm t_{\alpha/2}$ x *standard error of* $b_1$, which is *0.317+ 2.447* x *0.065 = 0.317* $\pm$ *0.159* or (*0.157, 0.477*).

2. Hypothesis testing that Ho : $\beta_2 = 0$

Use the following *t*–statistic: $t = \dfrac{b_2 - 0}{std - error, b_2}$

As $b_2$ = 0.433 and standard error $b_2$ = 0.033, *t*–statistic = 13.256 whilst $t_{0.5/2} = 2.447$ and the corresponding *p*-value = 0.00, it can be concluded that the null hypothesis Ho : $\beta_2 = 0$ is rejected.

One may be cautioned about conducting several *t*–tests on the *betas*. This is because if each *t*-test is conducted at $\alpha = 0.05$, the actual *alpha* that would cover all the tests simultaneously is considerably larger than 0.05.

For example, if significance tests are conducted on five betas, each at $\alpha = 0.05$, then if all the $\beta$ parameters (except $\beta_o$) are equal to zero, approximately $(1-0.95^2) = 0.227$ or 22.7% of the time we will incorrectly reject the null hypothesis at least once and conclude that some $\beta$ parameter differs from zero. If there are 10 betas, there are approximately 40% of the time we will make such mistake!

3. The adjusted coefficient of determination $R^2(adj)$

The coefficient of determination $R^2$ is also an important measure not to be ignored. The adjusted coefficient of determination $R^2(adj)$ has been adjusted to take into account the sample size and the number of independent variables. It is defined in terms of $R^2$ as follows:

$$R^2(adj) = 1 - \left[\frac{(n-1)}{n-(k+1)}\right](1-R^2)$$

In this drug experiments, we found $R^2$ = 0.971 and $R^2(adj)$ = 0.961 or 96.1%.

Do take note that the adjusted $R^2$ does not have the same interpretation as $R^2$. We know $R^2$ is a measure of goodness of fit, adjusted $R^2$ is instead a comparative measure of suitability of alternative nested sets of independent variables.