MA4605 Chemometrics – Lecture 3B Linear Models

We will break the conventional sequencing for this subject matter to look at multiple linear regression and some variable selection procedures in anticipation of next week's laboratory class.

Please be advised that past papers in the SULIS workspace. Next Monday we will look at 2009 sample paper.

Also be advised that some in-class exam papers similar to what should be expected for your in-class exam.

Confidence Intervals for Regression Coefficients

In the last class we looked how R can be used to determine the estimates and standard errors for the slope and intercept.

The following formulae can be used to compute the confidence intervals for both, for a specified significance level.

for significance level α, the confidence intervals are

$$(1-\alpha)\times 100\%\ \text{CI}\left[\widehat{\beta}_0\right] = \widehat{\beta}_0 \pm t_{n-2,1-\alpha/2}\text{SE}\left[\widehat{\beta}_0\right]\ ,$$
$$(1-\alpha)\times 100\%\ \text{CI}\left[\widehat{\beta}_1\right] = \widehat{\beta}_1 \pm t_{n-2,1-\alpha/2}\text{SE}\left[\widehat{\beta}_1\right]$$

These calculations provided the basis for end of semester examination questions in previous years, but that will not be the case for this year. To compute the confidence intervals for both estimates, we use the `confint()` command, specifying the name of the fitted model.

Recall the example used in the previous classes:

```
> Conc=c(0,2,4,6,8,10,12)
> Fluo=c(2.1,5.0,9.0,12.6,17.3,21.0,24.7)
>
> coef(Fit)
(Intercept)        Conc
   1.517857    1.930357

>
> Fit = lm(Fluo ~ Conc)
> confint(Fit)
                2.5 %   97.5 %
(Intercept) 0.75970 2.276014
Conc        1.82522 2.035495
```

Multiple Linear Regression

Previously we have seem SLR the case of one dependent variable Y explained by one independent variable X.

Multiple regression analysis is an extension of simple regression analysis, as described previously, to applications involving the use of two or more independent variables (predictors) to estimate the value of the dependent variable (response variable).

In the case of two independent variables, denoted by X_1 and X_2, the linear algebraic model is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i$$

The definitions of the above terms are equivalent to the definitions in previous classes for simple regression analysis, except that more than one independent variable is involved in the present case.

Based on sample data, the linear regression equation for the case of two independent variables is

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

The multiple regression equation identifies the best-fitting line based on ***the method of least squares***. In the case of multiple regression analysis, the best-fitting line is a line through n-dimensional space (3-dimensional in the case of two independent variables).

The calculations required for determining the values of the parameter estimates in a multiple regression equation and the associated standard error values are quite complex and generally involve matrix algebra. However, computer software, such as ***R***, is widely available for carrying out such calculations.

The assumptions of multiple linear regression analysis are similar to those of the simple case involving only one independent variable. For point estimation, the principal assumptions are that

(1) the dependent variable is a random variable,
(2) the relationship between the several independent variables and the one dependent variable is linear.

Additional assumptions for statistical inference (estimation or hypothesis testing) are that

(3) the variances of the conditional distributions of the dependent variable, given various combinations of values of the independent variables, are all equal,
(4) the conditional distributions of the dependent variable are normally distributed, and
(5) the observed values of the dependent variable are independent of each other. Violation of this assumption is called ***autocorrelation***,

Partial regression coefficient (or net regression coefficient). Each of the b_i regression coefficients is in fact a partial regression coefficient. A partial regression coefficient is the conditional coefficient given that one or more other independent variables (and their coefficients) are also included in the regression equation.

Conceptually, a partial regression coefficient represents the slope of the regression line between the independent variable of interest and the dependent variable given that the other independent variables are included in the model and are thereby statistically "held constant."

(Remark : We will refer to these values as the regression coefficients from now on, rather than as "slopes" . In the case of the "intercept estimate", we will just use the term "coefficient".)

Implementing a MLR model using R.

Implementing a MLR model in R is quite similar to fitting an SLR model. All one has to do is to specify the additional independent variables, using the following structure:

```
lm(y ~ x1 + x2 + …)
```

Example: Cheese Tasting

As an example, we shall use data on the taste of cheese, suggested in *Introduction to the Practice of Statistics* by D.S. Moore and G.P. McCabe, (Freeman, 1998).

The data give scores for the taste of a cheese (Taste) from 30 different formulations which caused variation in the concentration in the cheese of acetic acid (Acetic), hydrogen sulphide (H2S) and lactic acid (Lactic).

One would wish to model the dependence of the taste score on the concentrations of those three constituents, using the thirty observations

Case	Taste	Acetic	H_2S	Lactic
01	12.3	4.543	3.135	0.86
02	20.9	5.159	5.043	1.53
03	39.0	5.366	5.438	1.57
04	47.9	5.759	7.496	1.81
05	5.6	4.663	3.807	0.99
06	25.9	5.697	7.601	1.09
07	37.3	5.892	8.726	1.29
08	21.9	6.078	7.966	1.78
09	18.1	4.898	3.850	1.29
10	21.0	5.242	4.174	1.58
11	34.9	5.740	6.142	1.68
12	57.2	6.446	7.908	1.90
13	0.7	4.477	2.996	1.06
14	25.9	5.236	4.942	1.30
15	54.9	6.151	6.752	1.52
16	40.9	6.365	9.588	1.74
17	15.9	4.787	3.912	1.16
18	6.4	5.412	4.700	1.49
19	18.0	5.247	6.174	1.63
20	38.9	5.438	9.064	1.99
21	14.0	4.564	4.949	1.15
22	15.2	5.298	5.220	1.33
23	32.0	5.455	9.242	1.44
24	56.7	5.855	10.199	2.01
25	16.8	5.366	3.664	1.31
26	11.6	6.043	3.219	1.46
27	26.5	6.458	6.962	1.72
28	0.7	5.328	3.912	1.25
29	13.4	5.802	6.685	1.08
30	5.5	6.176	4.787	1.25

```
> FitAll

Call:
lm(formula = Taste ~ Acetic + H2S + Lactic, data = Cheese)

Coefficients:
(Intercept)       Acetic           H2S       Lactic
   -28.8768       0.3277       3.9118      19.6705
```

The fitted model is therefore

$$Taste^* = -28.9 + 0.33Acetic + 3.91H2S + 19.67Lactic$$

```
> FitAll = lm(Taste ~ Acetic + H2S + Lactic, data = Cheese)
> summary(FitAll)

Call:
lm(formula = Taste ~ Acetic + H2S + Lactic, data = Cheese)

Residuals:
    Min      1Q  Median      3Q     Max
-17.390  -6.612  -1.009   4.908  25.449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
Acetic        0.3277     4.4598   0.073  0.94198
H2S           3.9118     1.2484   3.133  0.00425 **
Lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared: 0.6518,     Adjusted R-squared: 0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

The Coefficient of Determination

The coefficient of determination, R^2, is a measure of the proportion of variability explained by, or due to the regression (linear relationship) in a sample of bivariate (i.e. X v Y) data. It is a number between zero and one and a value close to zero suggests a poor model.

A very high value of R^2 can arise even though the relationship between the two variables is non-linear. The fit of a model should never simply be judged from the R^2 value.

In the case of simple linear regression (i.e. bivariate data) the coefficient of determination is equivalent to the square of the correlation coefficient of X and Y.

In the case of MLR, the coefficient of determination is derived from Sums of Squares Identities (material we will cover soon).

The R^2 value is presented as part of the output of the summary command for a fitted model.

For the Cheese example : (Multiple) R^2 is found in the summary output

```
> FitAll = lm(Taste ~ Acetic + H2S + Lactic, data = Cheese)
> summary(FitAll)
…
…
…
Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared: 0.6518,     Adjusted R-squared: 0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

Overfitting

Overfitting describes the error which occurs when a fitted model is too closely fit to a limited set of observations. Overfitting the model generally takes the form of making an overly complex model (i.e. using an excessive amount of independent variables) to explain the behaviour in the data under study.

In reality, the data being studied often has some degree of error or random noise within it. Thus attempting to make the model conform too closely to sample data can undermine the model and reduce its predictive power.

(Remark: This will be the basis for a lab exercise)

Multicollinearity

Multicollinearity occurs when two or more independent in the model are highly correlated and, as a consequence, provide redundant information about the response when placed together in a model.

(Everyday examples of multicollinear independent variables are height and weight of a person, years of education and income, and assessed value and square footage of a home.)

From the Cheese example:

```
> cor(Cheese)

           Taste    Acetic       H2S    Lactic
Taste  1.0000000 0.5495393 0.7557523 0.7042362
Acetic 0.5495393 1.0000000 0.6179559 0.6037826
H2S    0.7557523 0.6179559 1.0000000 0.6448123
Lactic 0.7042362 0.6037826 0.6448123 1.0000000
```

Which independent variables have high correlation coefficients?

Consequences of high multicollinearity:

1. Increased standard error of estimates of the regression coefficients (i.e. decreased reliability of fitted model).
2. Often confusing and misleading results.

Adjusted R^2

Adjusted R^2 is used to compensate for the addition of independent variables to the model. As more independent variables are added to the regression model, unadjusted R^2 will generally increase but there will never be a decrease. This will occur even when the additional variables do little to help explain the dependent variable.

To compensate for this, adjusted R^2 is corrected for the number of independent variables in the model. The result is an adjusted R^2 than can go up or down depending on whether the addition of another variable adds or does not add to the explanatory power of the model. Adjusted R^2 will always be lower than unadjusted.

The adjusted R^2 is also presented in the output of the summary of a fitted model. It has become standard practice to report the adjusted R^2, especially when there are multiple models presented with varying numbers of independent variables.

For the Cheese example : Adjusted R^2 is found in the summary output

```
> FitAll = lm(Taste ~ Acetic + H2S + Lactic, data = Cheese)
> summary(FitAll)
…
…
…
Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared: 0.6518,     Adjusted R-squared: 0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

Variable Selection Procedures

Variable selection is intended to select the "best" subset of independent variables. Reasons for performing variable selections are:

- We want to explain the data in the simplest way. Redundant independent variables should be removed.
- [Rule of Thumb: Among several plausible regression models, the smallest model always fits the data best. The so-called "Law of Parsimony"]
- Unnecessary independent variables will reduce the precision in the (precise) estimation of other quantities that interested us.

- Multicollinearity is caused by having too many independent variables trying to do the same job. Removing excess predictors will aid interpretation.

Akaike's information criterion

The Akaike's information criterion (AIC), is a model selection metric. For a series of candidate fitted models, the model with a lowest AIC value is treated the best.

To compute the AIC for a candidate model in R, simply specify the name of the model as an argument to the `AIC()` function.

```
> AIC(FitAll)
[1] 229.7775
```

In next week's lab classes, we will use AIC and adjusted R2 to determine the best set of independent variables for fitting a (multiple) linear model.

Dummy Variables in Multiple Linear Regression

In regression analysis we sometimes need to modify the form of non-numeric variables, for example sex, or marital status, to allow their effects to be included in the regression model.

This can be done through the creation of dummy variables whose role it is to identify each level of the original variables separately.

Question 5

The quality of a certain pharmaceutical product is indicated by the percentage contamination of a by-product in the chemical synthesis. This by-product can be removed from the final batch, but only at considerable expense. It is thought that a cheaper way of improving quality would be to add an inhibitor, designed to stop the build-up of the by-product during the synthesis. Thirty two batches of the pharmaceutical product were produced at different pH settings, with and without the inhibitor. The percentage contamination of the final product was determined.

A dummy variable is used to represent the use of the inhibitor where

$$Ind = \begin{cases} 0 & \text{if no inhibitor has been used} \\ 1 & \text{if the inhibitor is used.} \end{cases}$$