MA4605 Chemometrics – Lecture 9A Linear Models

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

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

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

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
```

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.

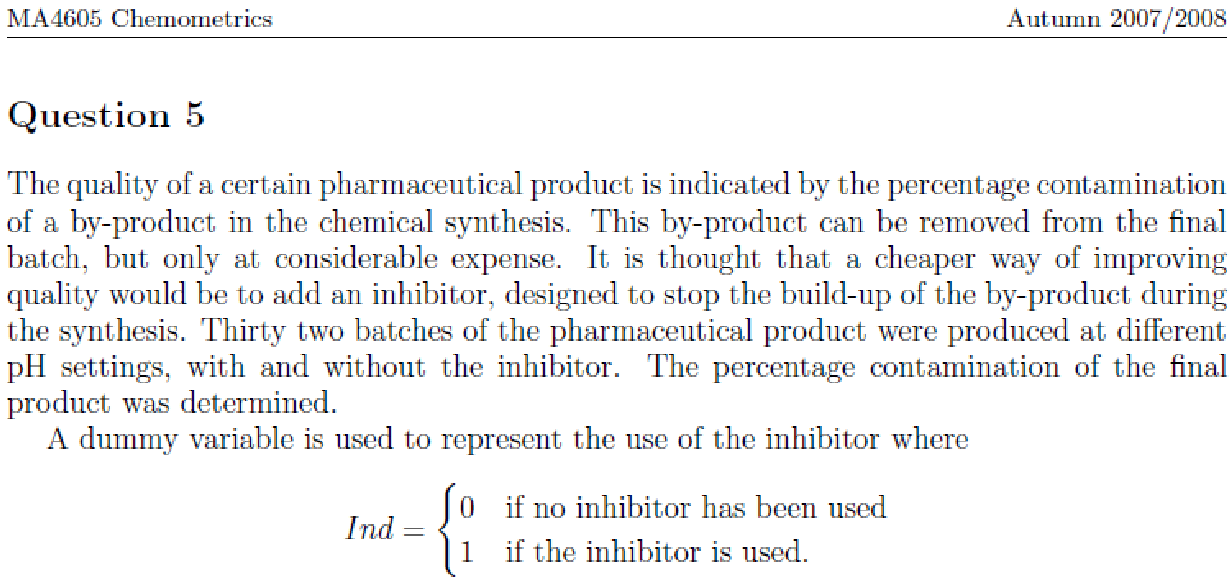
```
> AIC(FitAll)
[1] 229.7775
```

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.



MA4605 Chemometrics Autumn 2007/2008

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

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