

Multiple linear regression with Cheddar Cheese taste dataset
ST512 - Lab 4

1. Cheddar cheese data

- (a) Visit the page with the cheddar cheese taste data from the website lib.stat.cmu.edu/DASL/Datafiles/Cheese.html (link on ST512 website). Saving the data may not work because of all the `html` code, so cut and paste the data from the bottom of the page into the SAS program editor.
- (b) Using code similar to that below, write a program to read the data using an `INPUT` and a `CARDS` statement and write them to a temporary SAS dataset called 'cheese'. (If you prefer to avoid this exercise, you may borrow the code and data contained in the file "cheese.sas".)

```
DATA cheese;
  INPUT Case taste acetic h2s lactic;
  DROP case;
  CARDS;
1 12.3 4.543 3.135 0.86
2 20.9 5.159 5.043 1.53
...
```

- i. use `PROC PRINT; RUN;` to make sure the data were read in properly.
- ii. The command `DROP` gets rid of unwanted variables

- (c) Taste (y) is the response variable (y) of interest. Obtain the sample correlation coefficients between taste and the three independent variables, log acetic acid concentration, x_1 , log hydrogen sulfide x_2 and lactic acid concentration x_3 . (Independent variables have already been transformed in the data source.)

```
PROC CORR;  
  VAR taste acetic h2s lactic;
```

To get correlations, which reflect linear associations without standardization, try using the cov option:

```
PROC CORR COV;  
  VAR taste acetic h2s lactic;
```

Recall that correlations are always between -1 and 1, but the same is not true in general for covariances.

- i. True or false: there is evidence of a linear association between y and x_1 after averaging over observed values x_2 and x_3
- ii. True or false: there is evidence of a linear association between y and x_2 after averaging over observed values x_1 and x_3
- iii. True or false: there is evidence of a linear association between y and x_3 after averaging over observed values x_1 and x_2 .
- iv. True or false: there is a problem with multicollinearity here. Characterize the pairwise associations among the independent variables.

- (d) Get SAS/INSIGHT to produce a (4×4) scatterplot matrix of y, x_1, x_2, x_3 . An easier way to boot up SAS/INSIGHT than the click-sequence we've used is to just add the code

```
proc insight;run;
```

at the end of the current program. Then click-choose the library "WORK", then click-choose the appropriate dataset, then click the "OPEN" tab. Once the spreadsheet with the data appears, click **• Analyze • Scatter Plot (Y X)**. The order that variables are chosen will dictate which scatterplots go in which rows and columns in the scatterplot matrix. For this reason, make sure to click-highlight the response variable `taste` first. You can just choose them in the order they appear by click-dragging or by clicking, then pressing the SHIFT key, then clicking on the last variable on the list of variables on the left of the SAS: Scatter Plot (Y X). Then click the Y tab to select them for vertical axes. Repeat the process for the X tab. Click the OK tab.

- (e) What do the numbers in the diagonal boxes from the scatterplot matrix mean? (Refer to the output from PROC CORR to check your answer.)
- (f) You can fit the full multiple linear regression model MLR using all three variables in (at least) two ways:
- i. using SAS/INSIGHT
 - ii. using PROC REG

In SAS/INSIGHT, click **• Analyze • Fit (Y X)**, then defining `taste` as the response "Y" variable and all three of `acetic`, `h2s`, `lactic` as independent "X" variables. Before clicking OK, click Output and choose choose 95% C.I. for parameters. Click "Ok" until you get results.

To use PROC REG try code like the following:

```
PROC REG;  
  MODEL taste=acetic h2s lactic/CLB;  
RUN;
```

Use the output to address the following basic issues in MLR:

- i. What proportion of observed variation in taste is explained by the MLR model?
 - ii. What are the least squares estimates of the partial slopes for each independent variable?
 - iii. Observe the confidence intervals for these partial slopes.
 - iv. What is an estimate of the standard deviation of taste scores for cheeses with fixed values of the independent variables, x_1, x_2, x_3 .
- (g) See what happens when you consider discarding `acetic` and `lactic` using a `TEST` statement:

```
PROC REG;  
  MODEL taste=acetic h2s lactic;  
  TEST acetic=0, lactic=0;  
RUN;
```

Specify the nested and full models being compared here and report the results of the appropriate F ratio. Which model which you choose?

2. Obtain Mallows's C_p for each subset regression:

```
proc reg;  
  model y=x1-x3/selection=cp;  
run;
```

Which models are preferred under the criterion that $C_p \leq p+1$? Among these, which would you prefer? Why?

3. To see the difference, consider estimating the average taste response for cheese with $(x_1, x_2, x_3) = (5, 5, 1)$ with a 95% confidence interval. Note that even though $SS[E]$ is smaller and R^2 is bigger for the full model, the width of the 95% confidence interval for $\mu(x_1 = 5, x_2 = 5, x_3 = 1)$ is wider than under the more parsimonious model.