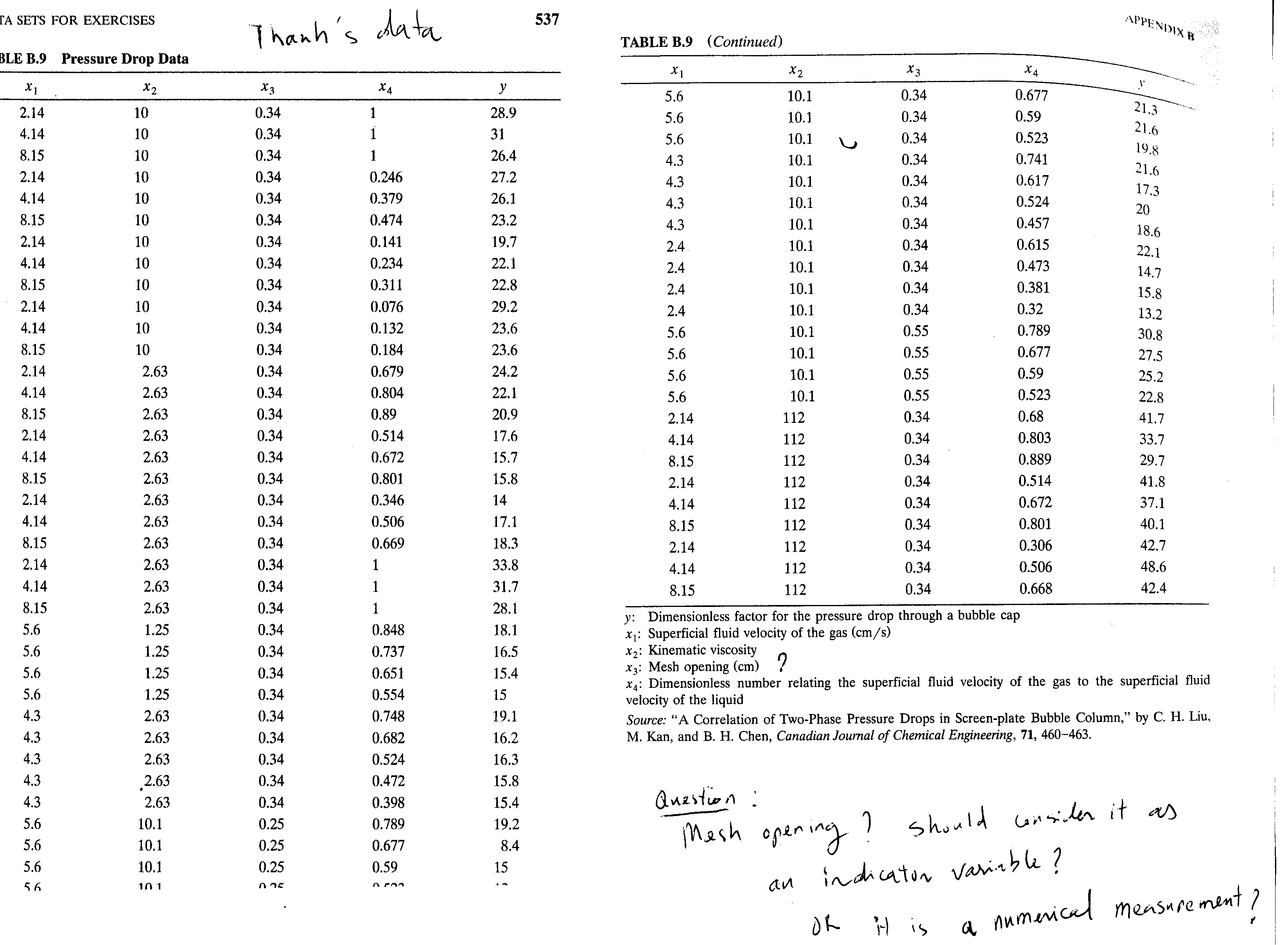
regression models – final project

**Thanh Doan – Student ID 0159701**

|  |  |
| --- | --- |
| Data set |  |



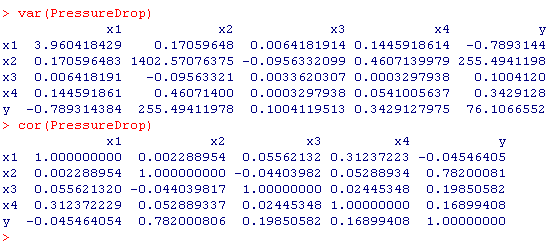
**Objective of the analysis**

* Analyzing the Pressure Drop Dataset to study if the response variable y (pressure drop through a bubble cap) is dependent on some or all of the repressors x1, x2, x3, x4.

**Assumption**

* All variables (y, x1, x2, x3, and x4) are numerical measurements. None of the repressors are categorical/indicator variables

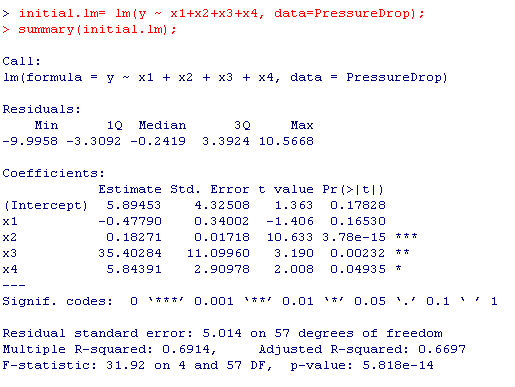
1. First I compute the variance/covariance of all study variables as well as Pearson Correlation to have some understanding about the variability of each variable and the pair-wise correlation between them. Though pair-wise correlation only tells us part of the story.



* **Comment:** Thevariance of x2 is very high. Most of the value of x2 are under 10 but then the last 9 data points (observation 54-62) has value 112 that makes the variability of x2 high.
* **Comment:**  The variance of x3 is very low, almost zero because most of data points has value x3=0.34. Some other values are different but not much different
* **Comment:** Among all repressors, x2 has highest correlation with the response variable y. Variables x1 and x2 are almost uncorrelated (Pearson correlation coefficient is 0.002) . Similarly x1 and y are almost uncorrelated (Pearson correlation coefficient is 0.045).

1. Fit and study an initial (full) multiple regression model

This initial model is used to study the residual analysis and understand **if** at least one regressor is important.



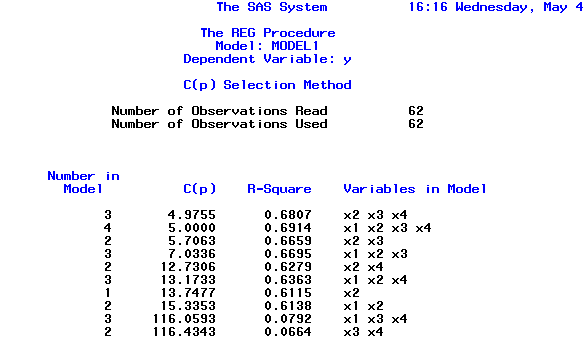
* **Comment:** The overall F test (F-statistic = 31.92) indicates that one or more regressors are important. The t-test in individual coefficients indicate that only x1 (superficial fluid velocity of the gas) is not important. We will analyze this further because this individual coefficient and its corresponding t-test only tell part of the story (and sometimes misleading) due to the nature of partial effect and if there is multi-collinearity or interaction in the data.

1. Plot of the residuals versus the predicted values and the normal probability plot of the residuals



* **Comment:** Thestudentized residuals are quite spreading evenly around zero. Most residuals are less than 3 std. deviation from the mean zero. There is no serious departure from normality.

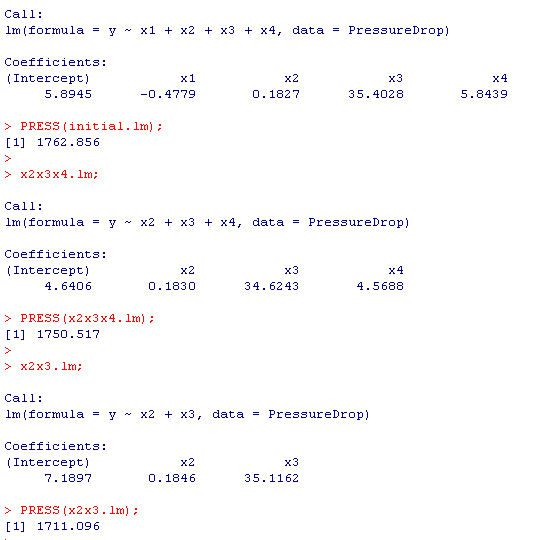
1. Since the number of regressors are small (4), we can use all-possible-regressions approach to fit all regression equations involving one candidate regressor, two candidate regressors, and so on using both adjusted R square and Mallows Cp statistics criteria to find the best subset.



**Comment:** Out of all possible regression models, three models below have similar small Cp statistics and similar adjusted R square values.

* The initial **full** model: y ~ x1+ x2+ x3 + x4
* The first reduced model: y ~ x2+ x3 + x4
* The second reduce model: y ~ x2+ x3

1. Perform validation of 3 regression models suggested at step 4.



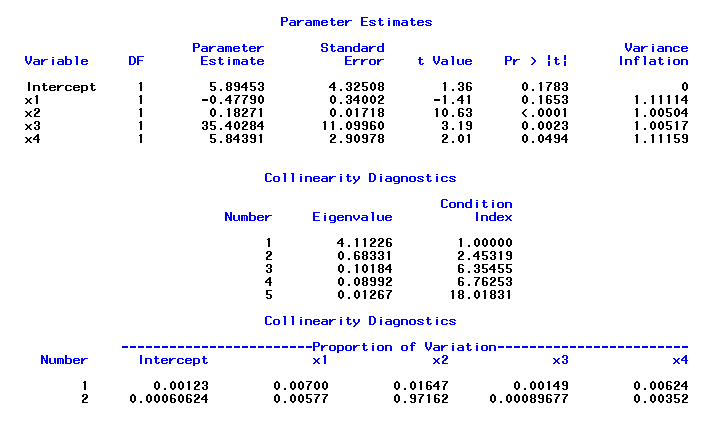
**Comment:**

* Here we use PRESS statistic to assess the predictive power of 3 suggested models. This is one form of data splitting in a sense that we use n-1 data points to estimate/train the model and use observation (i) to assess the model. The smaller the PRESS statistic the better.
* In the 3 model above… the last (most reduced) model
  + x2x3.lm = lm(y ~ x2+x3, data=PressureDrop);

has smallest PRESS statistic.

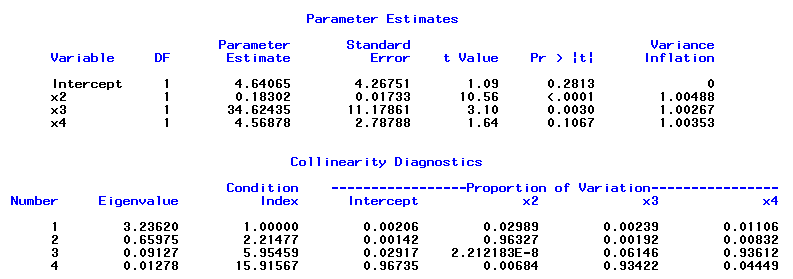
* This reduced model, y ~ x2+x3, also has small Cp value while it has equivalent adjusted R square value comparing to the initial full model.

1. Perform multi-collinearity diagnostics of 3 regression models suggested at step 4.



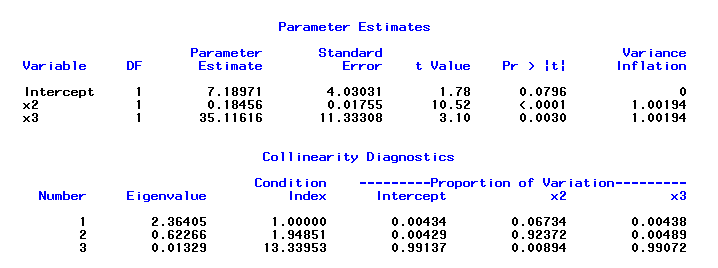
**Comment:**

* For the initial full model… there is no problem with collinearity in the data
* From VIF analysis… all variance inflation factors are around 1.0 and 1.1
* From Eigensystem analysis… all condition indices are less than 20 so there is no problem with collinearity in the set of 4 regressors



**Comment:**

* For the first reduced model y ~ x2+x3+x4, there is no problem with collinearity in the data
* From VIF analysis… all variance inflation factors are around 1.0
* From Eigensystem analysis… all condition indices are less than 16 so there is no problem with collinearity in the set of x2,x3,x4 regressors



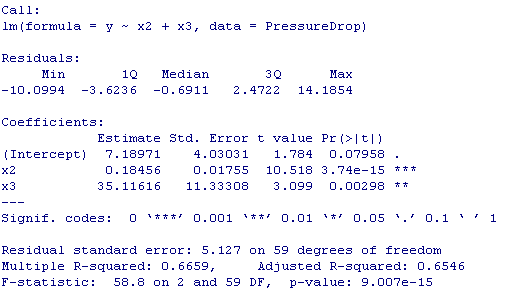
**Comment:**

* For the second reduced model y ~ x2+x3, there is no problem with collinearity in the data
* From VIF analysis… all variance inflation factors are around 1.0
* From Eigensystem analysis… all condition indices are less than 14. There is no problem with collinearity in the set of x2, x3 regressors

**Comments**

* After analysis in step 4, 5, 6 it is suggested that x2 and x3 are main explanatory variables of the response variable y (pressure drop through a bubble cap).
* In terms of adjusted R2 , both x2 and x3 collectively has the same explain power of the variability of y as all x1, x2, x3, x4 because their adjusted R2 values are similar
* Similarly in term of Cp statistic, the reduced model with only x2 and x3 in the regression equation is adequate.
* In terms of PRESS statistic, the reduced model with only x2 and x3 in the regression equation give the best PRESS statistic.
* If I have more time, I would two third of the data (randomly) for estimating the regression model and use 1/3 of the 62 data points to compute the root means square error (RMSE) or sum of square error to validate the model. However the PRESS statistic is also can be considered as a method to validate the model because it is equivalent to leave-one-out cross validation.
* In term of multi-collinearity issue, there is no problem in any model, including the full model

**With the above analysis I would suggest the reduced model with only x2 and x3 in the regression equation.**



1. Perform further analysis on the reduced model (only x2, x3 in the equation) to see if transformation or weighted regression is needed







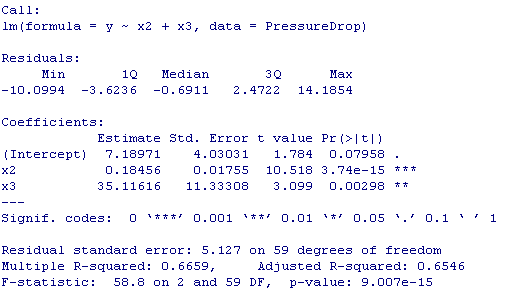


**Comments**

* From the cook distance plot there is no issue as the cook distance is less than 0.2 for all points
* From leverage analysis through the **H** matrix (see attached R/SPLUSs code) there is no serious problem.
* From the measure of influence statistics DFBETAS and DFFITS there is no serious influence point
* From the added-variable plot (in attached R code, but not shown here) and the component-residual plot (shown above) … it is suggestive that term X3 is not perfectly entered the model linearly.
* However, after different trying to transform y and/or x3 the resulted transformation does not give a better model in terms of adjusted R square as well as PRESS statistic.

**Recommended model**

* **y,** the response variable, is the pressure drop through a bubble cap
* x2 - Kinematic viscosity
* x3 - Mesh opening



* I don’t recommend 2 or three models in this analysis because this model has smaller subset of repressors, while having similar adjusted R square value and Cp statistic value.
* This model also has more predict power because its PRESS statistic is smaller comparing to two other models where x4 and x1 is included model.
* The conclusion of my analysis is similar to the author in the following article

<http://onlinelibrary.wiley.com/doi/10.1002/cjce.5450710317/abstract>

In terms of concluding that x2 and x3 are only explanatory variables.