Math 327, Chapter 3 Homework Part 2 - Coal liquifaction data

David Oniani

September 23, 2019

```r
# Edit the next line or otherwise ensure that the Appendix B data sets are loaded.
load ("./Appendix_B_data.Rdata")
```

```
## Warning: namespace 'emmeans' is not available and has been replaced
## by .GlobalEnv when processing object '.Last.ref_grid'
```

Appendix B, Table B.5, contains data on the Belle Ayr Liquifaction Runs.
Results of a kinetic study of thermal liquefaction of Belle Ayr coal are analyzed using a linear regression model (data from "(1978) Belle Ayr Liquefaction Runs with Solvent. Industrial Chemical Process Design Development, 17, 3"). One of the important performance measures is the production of CO_2 during the process. The process can be regulated with the help of several variables like total solvent(%), temperature (400, 425 or 450 centigrade) and hydrogen consumption(%). The variables are:

y = CO_2 (ppm)

x1 = Space time, min.

x2 = Temperaure, deg.C

x3 = Percent solvent (%)

x4 = Oil yield (g/100g MAF)

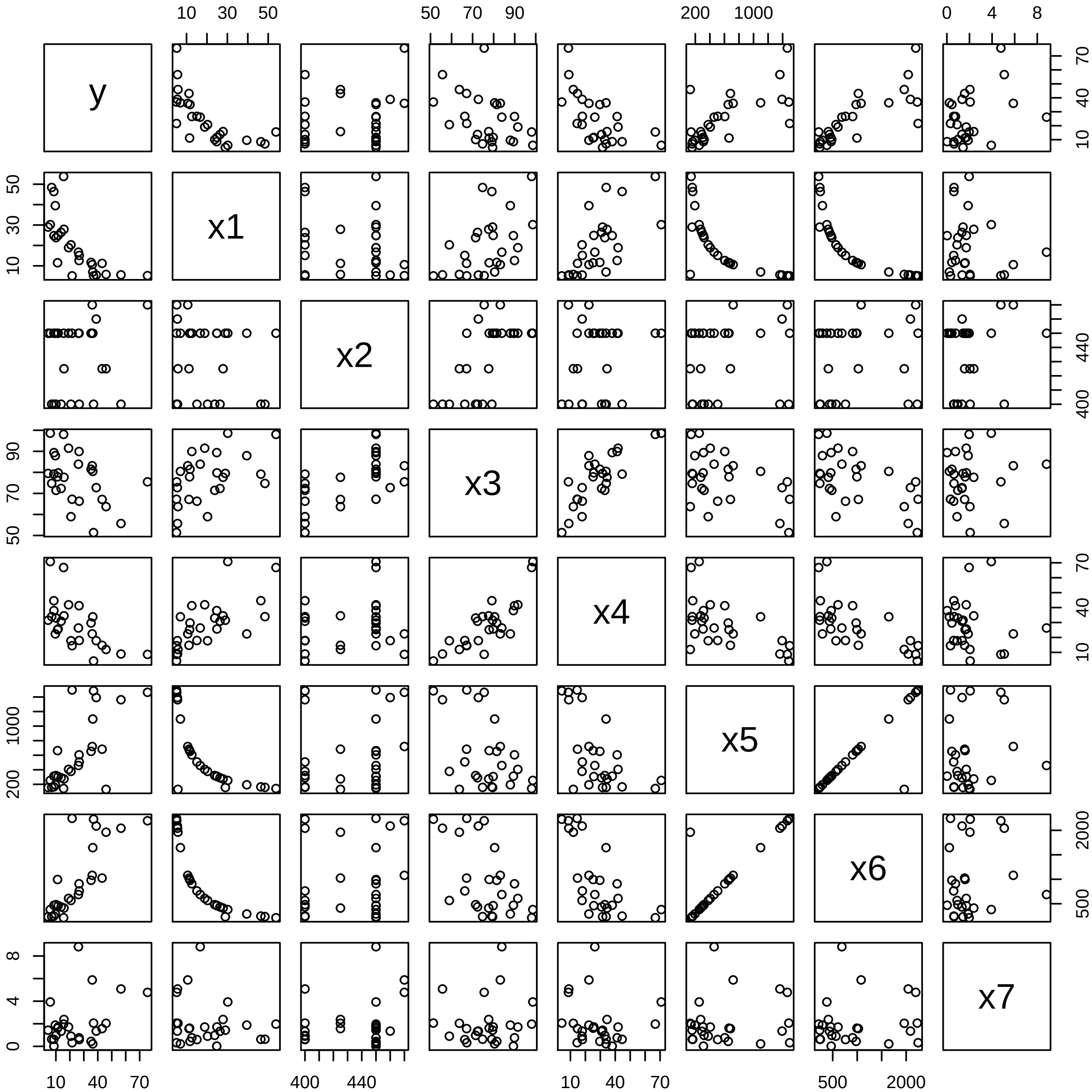
x5 = Coal total (%)

x6 = Solvent total (%)

x7 = Hydrogen consumption (%)

Produce a scatterplot matrix of all the data and fit the full first-order regression model.

```r
plot (dataB.5)
```




```r
fit1 = lm (y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data=dataB.5)
summary (fit1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = dataB.5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.035  -4.681  -1.144   4.072  21.214
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 53.937016  57.428952   0.939   0.3594
```

```
## x1          -0.127653   0.281498  -0.453   0.6553
## x2          -0.229179   0.232643  -0.985   0.3370
## x3           0.824853   0.765271   1.078   0.2946
## x4          -0.438222   0.358551  -1.222   0.2366
## x5          -0.001937   0.009654  -0.201   0.8431
## x6           0.019886   0.008088   2.459   0.0237 *
## x7           1.993486   1.089701   1.829   0.0831 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.61 on 19 degrees of freedom
## Multiple R-squared:  0.728,  Adjusted R-squared:  0.6278
## F-statistic: 7.264 on 7 and 19 DF,  p-value: 0.0002674
```

Remove the least significant variable and refit

```r
# Add R code here to refit the model
# We removed x5 as it was the least significant variable
fit2 = lm (y ~ x1 + x2 + x3 + x4 + x6 + x7, data=dataB.5)
summary (fit2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x6 + x7, data = dataB.5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.321  -4.703  -1.152   3.918  20.956
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 54.955652  55.814729   0.985  0.33658
## x1          -0.126328   0.274585  -0.460  0.65043
## x2          -0.232979   0.226239  -1.030  0.31540
## x3           0.832064   0.745859   1.116  0.27783
## x4          -0.441126   0.349558  -1.262  0.22148
## x6           0.018830   0.005995   3.141  0.00514 **
## x7           1.989598   1.063066   1.872  0.07597 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.36 on 20 degrees of freedom
## Multiple R-squared:  0.7274, Adjusted R-squared:  0.6456
## F-statistic: 8.895 on 6 and 20 DF,  p-value: 8.468e-05
```

Remove the next least significant variable and refit

```r
# Add R code here to refit the model
# We removed x1 as it was the least significant variable
```

```r
fit3 = lm (y ~ x2 + x3 + x4 + x6 + x7, data=dataB.5)
summary (fit3)
```

```
##
## Call:
## lm(formula = y ~ x2 + x3 + x4 + x6 + x7, data = dataB.5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.0541  -4.0883  -0.6269   4.4727  19.9486
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.664125  50.166956   0.890 0.383386
## x2          -0.209230   0.216097  -0.968 0.343952
## x3           0.789045   0.725953   1.087 0.289396
## x4          -0.459129   0.340778  -1.347 0.192244
## x6           0.020095   0.005226   3.845 0.000941 ***
## x7           2.011910   1.041835   1.931 0.067078 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.16 on 21 degrees of freedom
## Multiple R-squared:  0.7245, Adjusted R-squared:  0.6589
## F-statistic: 11.05 on 5 and 21 DF,  p-value: 2.592e-05
```

Continue until Adjusted R^2 is maximized

```r
# Add R code here
# We removed x2 as it was the least significant variable
fit4 = lm (y ~ x3 + x4 + x6 + x7, data=dataB.5)
summary (fit4)
```

```
##
## Call:
## lm(formula = y ~ x3 + x4 + x6 + x7, data = dataB.5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7958  -5.8786   0.3351   5.6663  20.4730
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.133571  19.363209  -0.007 0.994558
## x3           0.149407   0.300528   0.497 0.624017
## x4          -0.231610   0.246452  -0.940 0.357535
## x6           0.016570   0.003745   4.425 0.000214 ***
## x7           2.011550   1.040353   1.934 0.066141 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.15 on 22 degrees of freedom
```

```
## Multiple R-squared:  0.7122, Adjusted R-squared:  0.6599
## F-statistic: 13.61 on 4 and 22 DF,  p-value: 9.905e-06
```

```r
# We removed x3 as it was the least significant variable
fit5 = lm (y ~ x4 + x6 + x7, data=dataB.5)
summary (fit5)
```

```
##
## Call:
## lm(formula = y ~ x4 + x6 + x7, data = dataB.5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.5535  -4.6608   0.1209   4.8798  21.7928
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.612021   7.957872   1.082 0.290379
## x4          -0.141944   0.165180  -0.859 0.399029
## x6           0.016440   0.003674   4.475 0.000172 ***
## x7           2.162891   0.978402   2.211 0.037281 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.979 on 23 degrees of freedom
## Multiple R-squared:  0.709,  Adjusted R-squared:  0.671
## F-statistic: 18.68 on 3 and 23 DF,  p-value: 2.311e-06
```

```r
# We removed x4 as it was the least significant variable
# Continuing this process will result in a significant decrease of R^2, thus we stop here
fit6 = lm (y ~ x6 + x7, data=dataB.5)
summary (fit6)
```

```
##
## Call:
## lm(formula = y ~ x6 + x7, data = dataB.5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.2035  -4.3713   0.2513   4.9339  21.9682
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.526460   3.610055   0.700   0.4908
## x6          0.018522   0.002747   6.742 5.66e-07 ***
## x7          2.185753   0.972696   2.247   0.0341 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.924 on 24 degrees of freedom
## Multiple R-squared:  0.6996, Adjusted R-squared:  0.6746
## F-statistic: 27.95 on 2 and 24 DF,  p-value: 5.391e-07
```

```r
# Some additional statistics in the form of confidence intervals
confint (fit6)
```

```
##                    2.5 %     97.5 %
## (Intercept) -4.92432697 9.97724714
## x6           0.01285196 0.02419204
## x7           0.17820756 4.19329833
```

Interpret the the final model (parameter estimates, adjusted R^2, residual standard error)

The mean response changes between 0.013 and 0.024 C02 (ppm) per Solvent total (x_6), for any 1-unit increase in the predictor with 95% confidence, holding all other predictors fixed.

The mean response changes between 0.178 and 4.193 C02 (ppm) per Hydrogen consumption (%) (x_7), for any 1-unit increase in the predictor with 95% confidence, holding all other predictors fixed.

It should be noted that both estimates Solvent total (x_6) and Hydrogen consumption (%) (x_7) are statistically significant and have the p-values of $5.66 * 10^{-7}$ and 0.0341 respectively.

The coefficient of determination (Adjusted R^2) is 0.6746 which means that 67.46% of variation in CO2 (ppm) is explained by the regression model (in this case, two regressor variables x_6 and x_7).

The residual standard error is 9.924 which tells us that, on average, our predictions are 9.924 CO2 (ppm) off from the real value.

Second model-building exercise

Since the scatterplot of Y vs X1 is curved, an X1 quadratic predictor is added to the model. Using that model, repeat the model-building procudure as above.

```r
# Add x1 squared predictor to the data frame
dataB.5$x1sq = (dataB.5$x1 - mean (dataB.5$x1))^2
fit1A = lm (y ~ x1 + x1sq + x2 + x3 + x4 + x5 + x6 + x7, data=dataB.5)
summary (fit1A)
```

```
## 
## Call:
## lm(formula = y ~ x1 + x1sq + x2 + x3 + x4 + x5 + x6 + x7, data = dataB.5)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -17.5237  -2.9468  -0.0001   3.1214  23.6842 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 89.333148  57.394545   1.556   0.1370  
## x1          -1.219570   0.648007  -1.882   0.0761 .
## x1sq         0.036565   0.019799   1.847   0.0813 .
## x2          -0.229474   0.219156  -1.047   0.3089  
## x3           0.774345   0.721424   1.073   0.2973  
## x4          -0.462944   0.338030  -1.370   0.1877  
## x5          -0.001453   0.009099  -0.160   0.8749  
## x6           0.003049   0.011882   0.257   0.8004  
## x7           2.017500   1.026610   1.965   0.0650 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.999 on 18 degrees of freedom
## Multiple R-squared:  0.7713, Adjusted R-squared:  0.6697
## F-statistic: 7.589 on 8 and 18 DF,  p-value: 0.0001889
```

Repeat the predictor removal process as used above. Interpret the final model you obtain and compare that model to the final model you obtained above.

```r
# We removed x5 as it was the least significant variable
fit2A = lm (y ~ x1 + x1sq + x2 + x3 + x4 + x6 + x7, data=dataB.5)
summary (fit2A)
```

```
##
## Call:
## lm(formula = y ~ x1 + x1sq + x2 + x3 + x4 + x6 + x7, data = dataB.5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.565  -2.976   0.585   3.063  23.497
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 90.184582  55.661423   1.620   0.1217
## x1          -1.221300   0.631082  -1.935   0.0680 .
## x1sq         0.036656   0.019277   1.902   0.0725 .
## x2          -0.232322   0.212753  -1.092   0.2885
## x3           0.779622   0.701942   1.111   0.2806
## x4          -0.465181   0.328964  -1.414   0.1735
## x6           0.002216   0.010398   0.213   0.8335
## x7           2.014647   0.999785   2.015   0.0583 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.74 on 19 degrees of freedom
## Multiple R-squared:  0.771,  Adjusted R-squared:  0.6866
## F-statistic: 9.138 on 7 and 19 DF,  p-value: 5.909e-05
```

```r
# We removed x6 as it was the least significant variable
fit3A = lm (y ~ x1 + x1sq + x2 + x3 + x4 + x7, data=dataB.5)
summary (fit3A)
```

```
##
## Call:
## lm(formula = y ~ x1 + x1sq + x2 + x3 + x4 + x7, data = dataB.5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1919  -2.7362   0.3213   2.8169  24.0502
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 91.99431   53.68124   1.714 0.102040
## x1          -1.33812    0.30534  -4.382 0.000288 ***
## x1sq         0.04011    0.01020   3.933 0.000823 ***
## x2          -0.22002    0.19982  -1.101 0.283948
```

```
## x3           0.73256    0.65022   1.127 0.273233
## x4          -0.45631    0.31844  -1.433 0.167317
## x7           2.02227    0.97501   2.074 0.051196 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.504 on 20 degrees of freedom
## Multiple R-squared:  0.7704, Adjusted R-squared:  0.7016
## F-statistic: 11.19 on 6 and 20 DF,  p-value: 1.681e-05
```

```r
# We removed x2 as it was the least significant variable
fit4A = lm (y ~ x1 + x1sq + x3 + x4 + x7, data=dataB.5)
summary (fit4A)
```

```
##
## Call:
## lm(formula = y ~ x1 + x1sq + x3 + x4 + x7, data = dataB.5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.815  -3.466   1.806   3.778  23.306
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.892670  16.982063   2.114  0.04669 *
## x1          -1.107348   0.223170  -4.962 6.56e-05 ***
## x1sq         0.035236   0.009235   3.816  0.00101 **
## x3           0.088720   0.285781   0.310  0.75928
## x4          -0.233494   0.247108  -0.945  0.35545
## x7           2.022689   0.979927   2.064  0.05158 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.552 on 21 degrees of freedom
## Multiple R-squared:  0.7565, Adjusted R-squared:  0.6986
## F-statistic: 13.05 on 5 and 21 DF,  p-value: 7.504e-06
```

```r
# We removed x3 as it was the least significant variable
# Continuing this process will result in a significant decrease of R^2, thus we stop here
fit5A = lm (y ~ x1 + x1sq + x4 + x7, data=dataB.5)
summary (fit5A)
```

```
##
## Call:
## lm(formula = y ~ x1 + x1sq + x4 + x7, data = dataB.5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.716  -3.393   1.387   3.379  24.102
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.972069   4.454515   9.198 5.40e-09 ***
## x1          -1.108260   0.218520  -5.072 4.43e-05 ***
## x1sq         0.034894   0.008979   3.886 0.000795 ***
## x4          -0.175453   0.158229  -1.109 0.279468
```

```
## x7           2.106617   0.922351   2.284 0.032387 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.354 on 22 degrees of freedom
## Multiple R-squared:  0.7554, Adjusted R-squared:  0.7109
## F-statistic: 16.99 on 4 and 22 DF,  p-value: 1.745e-06
```

```
# Some additional statistics in the form of confidence intervals
confint (fit5A)
```

```
##                   2.5 %      97.5 %
## (Intercept) 31.73397110 50.21016744
## x1          -1.56144163 -0.65507798
## x1sq         0.01627347  0.05351432
## x4          -0.50359935  0.15269391
## x7           0.19377877  4.01945568
```

The mean response changes between -1.561 and -0.655 C02 (ppm) per Space time (min) (x_1), for any 1-unit increase in the predictor with 95% confidence, holding all other predictors fixed.

The mean response changes between -0.504 and 0.152 C02 (ppm) per Oil yield (g/100g MAF) (x_4), for any 1-unit increase in the predictor with 95% confidence, holding all other predictors fixed.

The mean response changes between 0.194 and 4.019 C02 (ppm) per Hydrogen consumption (%) (x_7), for any 1-unit increase in the predictor with 95% confidence, holding all other predictors fixed.

The mean response changes between 0.016 and 0.054 C02 (ppm) per Space time (min) (x_1^2), for any 1-unit increase in the predictor with 95% confidence, holding all other predictors fixed.

It should be noted that out of all estimates, only Oil yield (g/100g MAF) (x_4) is not statistically significant with the p-value of 0.279. Estimates $x1$, $x1sq$, and $x7$ are all statistically significant with the p-values equal to $4.43 * 10^{-5}$, 0.000795, and 0.0324 respectively.

The coefficient of determination (Adjusted R^2) is 0.7554 which means that 75.54% of variation in CO2 (ppm) is explained by the regression model (in this case, three regressor variables x_1, x_4, x_7, and the quadratic term x_{1sq}).

The residual standard error is 9.354 which tells us that, on average, our predictions are 9.354 CO2 (ppm) off from the real value.

first model

The coefficient of determination (Adjusted R^2) is 0.6746 which means that 67.46% of variation in CO2 (ppm) is explained by the regression model (in this case, two regressor variables x_6 and x_7).

The residual standard error is 9.924 which tells us that, on average, our predictions are 9.924 CO2 (ppm) off from the real value.

When compared wit the first model, the second model seems to be better. Introducing quadratic term both increased Adjusted R^2 and decreased residual standard error. Adjusted R^2 value went up from 0.6746 to 0.7554 (which means being, on average, roughly 8% improvement). Residual standard error went down from 9.924 to 9.354 which tells us that, on average, we are 0.57 CO2 (ppm) more accurate in our predictions. Hence, the second model with the quadratic term is overall a better model with higher Adjusted R^2 value and lower residual standard error.