Heating up Coal: An Analysis on Higher Heating Values of Coal

Last name: Xiang
First name: Yi
Course section: STA302H1F-Summer 2017

This a linear regression analysis report on the effects of moisture, volatile matter, ash, and fixed carbon has on low rank coal. The data is from a dataset called coal_hhv.dat from the University of Florida's statistics department. The data was used for a science journal about Fuel Processing Technology. In the research, the author used Non-Linear Regression to analyse the data to determine the relationship between Higher Heating Value of coal (in MJ/kg) to other possible predictors. I was interested in the data originally because of I was curious to learn more about Coal Energy.

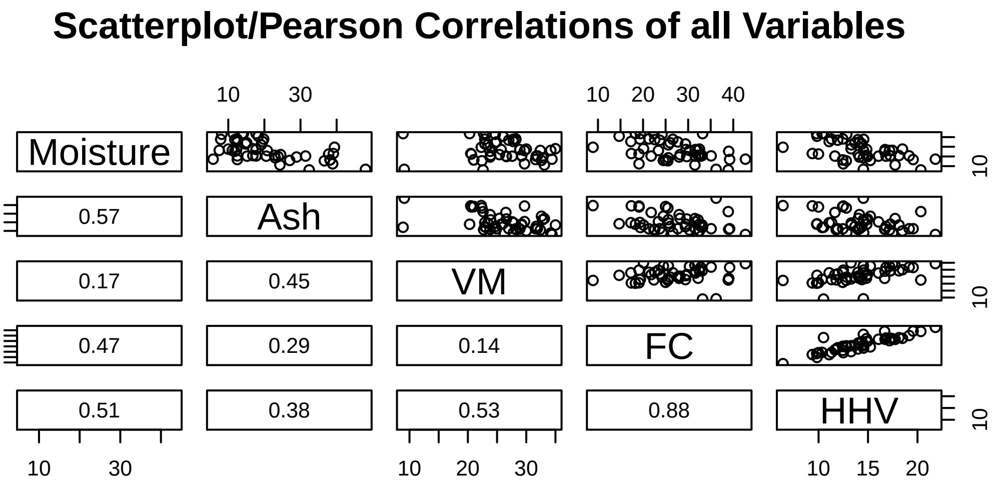
The dataset contains 6 variables:

1. Tested Coal - The specific low grade coal that was tested.
2. Higher Heating Value - The energy potential (in MJ/kg) for the coal.
3. Ash - The amount of Ash (in weight % in relation with the coal tested).
4. Volatile Material - The amount of Volatile Material (in weight %).
5. Moisture - The amount of Moisture (in weight %).
6. Fixed Carbon - The amount of Carbon (in weight %).

I choose to look specifically at how the predictors affect the Heating Values of Coal (denounced "HHV"), because of primarily 2 reasons. Firstly, the data is from another study that was looking at the same thing, albeit with another method. The second being it was the variable that showed the strongest level of correlation with all other 4 variables when looking at the Pearson correlations of all variables. This is why I've chosen HHV to be my response variable and Ash, Volatile Material (denounced "VM"), Moisture, Fixed Carbon (denounced "FC") to be my explanatory variables.

Looking purely at the context of the data, I expect a linear correlation between what is essentially, the relationship between the heating point of coal and the components of the coal. It makes logical sense that the kinetic energy potential of a material is affected by it's materials. The study that this data was gathered from used a nonlinear regression model to predict HHV. I am curious to see if I can create a linear regression model for the data.

Data Analysis



There is strong correlation between HHV and FC, and weaker correlation with Moisture, VM, and especially Ash. As noted in the introduction.

Testing with a model with all 4 variables

```
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 10.802342  84.331896  0.128093 0.898659
## Moisture    -0.141251   0.845482 -0.167066 0.868083
## Ash         -0.118289   0.840421 -0.140750 0.888710
## VM           0.119806   0.842657  0.142176 0.887590
## FC           0.239034   0.843167  0.283495 0.778127
```

```
## [1] 0.949229
```

From the summary alone, this model is not usable in any way due to the fact that it suggests that all models show no evidence that the coefficients of all predictory variables are non-zero. The high adjusted R-Squared value implies that the model is a bad fit, implying we are over fitting our model.

Preforming backwards Stepwise regression (see Appendix) tells us to remove variable “Ash” from our current model (This matches with our correlation plot, as it showed the low correlation between Ash and HHV). Doing so produces.

```
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) -1.066756   0.821116 -1.299154 0.200508
## Moisture    -0.022261   0.011490 -1.937457 0.058981
## VM           0.238378   0.019026 12.529165 0.000000
## FC           0.357686   0.016452 21.741329 0.000000
```

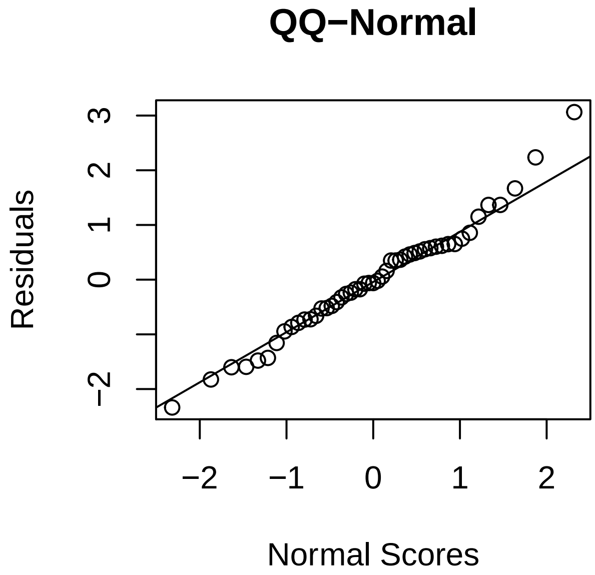
```
## [1] 0.950335
```

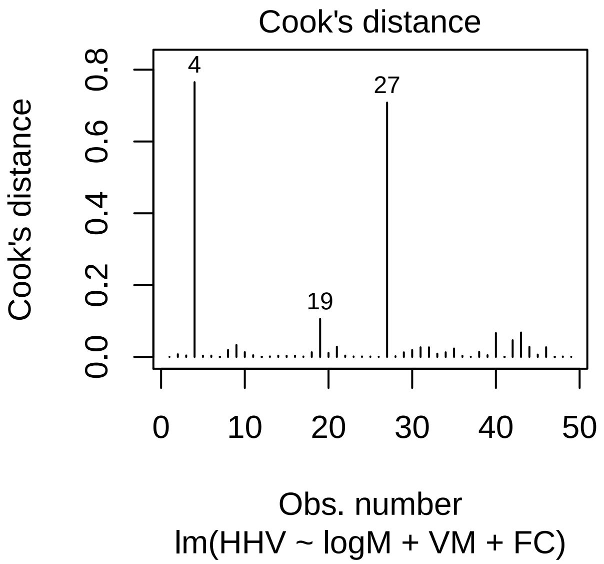
Our second fit produces much more palatable results. Offering strong proof for the coefficients of VM and FC to be non-zero. There is weak evidence for the coefficient of Moisture to be non-zero. We can transform the Moisture variable with log-transformations since the errors are not homoscedasticity (See Appendix)

```
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept)  0.656327   1.117520  0.587307 0.559933
## logM        -0.720626   0.246144 -2.927654 0.005341
## VM           0.245077   0.018037 13.587115 0.000000
## FC           0.350665   0.015793 22.203656 0.000000
```

```
## [1] 0.954801
```

Now we have much stronger evidence to believe that the coefficient of *log*Moisure is non-zero. Lets take a look at the residuals of this model, in particular the qqplot.





The QQ graph and the sharpio-wilks test (See Appendix) shows that residuals are non-normal. However the Cook's distance graph shows that there are 2 outliers in the standarized residuals that greately influence the normality of the errors in this model. Because of our already very limited sample size of n=50, it is unjustified to remove the data samples that cause these outliers. Looking at the qqplot, these outliers are not an overall problem to the general normality of the model.