REPORT:

In this problem we are given D(Cu) and Fe-Domains, and our job is to find a relation between these two for different elements.

The relation between these two will vary depending upon the nature of elements and the nature of the system. System here refer to Fe-Ni-S, Fe-Ni-C, Fe-S-C , etc.

* S1: We have an initial model, which is derived from equations of thermodynamics, but still we will first try to come up with an equation, on the basis of data collected.
  + Plotted Fe-Domains vs Dinv
    - Looks like hyperbola, quadratic, or linear if we remove a few points.
  + FeD/Dinv (A1)
    - Linear: r-squared = 0.546
    - Little pattern in residuals
  + FeD/Dinv (A2)
    - Squared: r-squared = 0.785
    - Non-linearity(NL) in residuals ~ maybe due to 3 outliers
    - Heteroscedasticity(HS) present
    - To remove HS, applied log transformation on Y
  + FeD/Ln(Dinv) (A3)
    - Squared : r-squared = 0.716
    - Quite a good fit
    - HS removed
    - NL in residuals not visible
    - But the problem with quadratic fit is that it has a maximum or a minimum, so after going in a particular direction, it will change its direction. We can always crop one side of the extremum, but cropping won’t always help.
  + FeD/Dinv (A4)
    - Square-root : r-squared = 0.84
    - Clear trend in residuals
  + Ln(FeD)/Dinv (A5)
    - Log : r-squared = 0.721
    - Clearly missing high leverage points (Fe-S-C)
    - High p-value of the intercept – 0.40
    - Residuals show HS, and outliers
    - Applying log transformation to deal with HS
  + Ln(FeD)/Ln(Dinv) (A6)
    - Log – 0.67
    - HS reduced, what is left may be due to fewer points on the left side
    - R-squared decreased in this case but the fit improved, Fe-S-C points don’t look like outliers in this case.
    - But p-value is still very high (0.363), may be due to high leverage points.
    - A5 vs A6: p-value=insignificant; HS=reduced; Outliers=accommodated
  + FeD/Dinv (A7)
    - Hyperbola – 0.865, r-squared quite high
    - Issues – under-fits for 2<x<5, over-fits for x>5
    - HS present, in fewer than 20% points, applying transformation
  + FeD/Ln(Dinv) (A8)
    - Hyperbola – 0.654, r-squared drastically reduced
    - HS still present
    - Non-linear behavior seen in residuals – under-fits for x<4, over-fits for x>4
* S2: Until now all the models which were able to fit the data, had high p-values. To deal with this problem we will fit all the models again, after removal of Outliers/High leverage points. These points mostly belonged to Fe-S-C systems.
  + FeD/Dinv (A10)
    - Linear – 0.608
    - Non-linearity clearly seen in residuals
  + FeD/Ln(Dinv) (A11)
    - Linear – 0.491
    - R-squared unacceptable
    - Non-linearity still present in residuals.
  + FeD/Ln(Dinv)
    - Squared – 0.600
    - R-squared reduced, but non-linearity not seen.
  + FeD/ Dinv
    - Squared – 0.736
    - NL present
  + Ln(FeD)/Dinv
    - Log – 0.682
    - Residuals show trends, HS present, transformation applied
  + Ln(FeD)/Ln(Dinv)
    - Log – 0.543
    - R-squared reduced
    - Mojor concern – p-value= 0.239
  + FeD/Dinv
    - Hyperbola – 0.695
    - NL,HS present
  + FeD/Ln(Dinv)
    - Hyperbola – 0.537
    - NL, r-squared
    - Looks like hyperbola fitted well only with outliers
    - Hyperbola rejected

S1:

Model A6 was fitting the data well, but had a high p-value, initial approach was to remove the high leverage points and see if the p-value decreases.

Another model A7 (hyperbola) has a very high r-squared statistic, but the model doesn’t fit the data well as non-linear behavior could be clearly seen in the data.

S2:

Model A7 and A8 (hyperbola) which showed promise are rejected, as there r-squared has now reduced significantly, and NL from the residuals is not that much reduced.

Model A6, after removing the points also shows the same problem, now the r-squared is even lower, and the p-value is (0.236).

CONCLUSION1: If we are only considering 1 feature which is Fe Domains, then the best model seems to be Ln(FeDomains) vs Ln(Dinv), which is also obtained from the thermodynamic equations, but the high p-value associated with this model needs to be checked.

* S3: Until now the major cause of trouble were the Fe-S-C points, also all the models until now were fitted on the Cu dataset only. In the journals, it is mentioned that for Siderophile elements only Fe Domains are enough to predict the segregation coefficient of a trace constituent, even if exact Fe/Ni ratio, and temperature are not known.
* So now we are going to check the effect of Fe-S-C points on the Siderophile elements.
  + Here we assumed that in case of Siderophile elements only the value of Fe Domains are important, and not which point belonged to which system. So if we remove a few points less than 10% from a specific system (Fe-S-C), then there will be almost no change in the fit.
  + On fitting the models on Co, Mo, Os, Pt, Au, Re, Ru it is seen that there is clear effect on the fit of the model on removing points from Fe-S-C system. There were very few points in case of Mn, Rh to make any comment.
  + So based on above exercise it can be concluded that Fe/Ni ratio plays a role in deciding the segregation co-efficient of trace elements in case of Siderophile elements also.
  + One reason which may be the cause of error in the above observation is that almost all the Fe-S-C points are high leverage points ( i.e. all these points have unusual value of Fe Domains), so removing these points have a significant impact on the model. Conducting more experiments with values of Fe-Domains in the region where Fe-S-C belong may solve this problem
* S4: Originally I was asked to increase the order of the regression so as to get a better fit of the data. On applying linear regression of order 2, to various models previously done by Soyyeb Sir, the r-squared statistic improved significantly in the models, and a much better fit was obtained but we also obtained very high p-values. So increasing the order we will get a better fit but the reliability of these model will be significantly decreased.
* P-Values: Roughly speaking, we interpret p-values as follows: a small p-value indicates that it is unlikely to observe to such a substantial association between the predictor and the response due to chance, in the absence of any real association between the predictor and the response. Hence if we see a high p-value we can infer that the association between the predictor and response may be due to chance.
* In S4: This is seen as a general trend that all the higher order model are very easily affected by high leverage points, and so we get a higher p-value.

Further we can check for trends in different systems separately, or try to add qualitative predictors (in form of different systems) to the fit.

Also due to the various assumptions in parametric methods like linear regression we are facing challenges, so we can try some non-parametric methods also.