Regression Analysis

June 16, 2015

1 Introduction

Regression analysis is a powerful technique that can be used to address various research questions. In this report, we are going to use it to check how q levels are affected by cu and phi. In particular, the type of regression we are going to use is "Multiple Linear Regression". Linear regression is the process of finding the best-fitting straight line through data points (this line is sometimes referred to as the regression line). Multiple means we have more than one input variable (also known as predictor), hence, we are trying to fit a plane or hyper-plane rather than a line. The input variables in our case are cu, phi and as. Linear means that we are trying to find a combination of the input variables such that each variable is multiplied by a coefficient and then we sum the products. The idea is to use this linear combination of input variables to model their relationship with an output variable (in our case, this is q).

2 The Modelling Tool

In order to fit models, we are going to use R [**?**] which is a powerful and easy to use tool for statistical computing and graphics. R makes it easy to manipulate data and perform calculations as well as display information graphically. It also facilitates modelling (linear and nonlinear) and other statistical processes.

3 The Research Question

Before we emark on any analysis, let us be clear about what we are addressing.
In this report, we will be addressing whether there is any association between VARIABLE(s) and Q.

4 Diagnostics for Examining a Fit

There are several diagnostics that can be used to explore the goodness of fit of a model. In the remainder of this report (Section 11) we are going to use the following:

4.1 R-squared

This value calculates the percentage of variation of the output explained by the input variables in the model. This means the higher the value of R-squared the better the model.

4.2 R-squared adjusted

This value is similar to R-squared but it accounts for the number of input variables in the model, hence, it is sometimes preferred to R-squared.

4.3 Residuals

A residual is the difference between the actual value and predicted value for each point (or record) in the data. Histograms are often used to check the distribution of residuals. Also, they are plotted against each input variable. If a model fits well, the residuals will be small and will be no pattern of their distrbution around zero (i.e. they should be evenly spread around zero).

4.4 Deviance

The deviance is a statistic that is used to determine the quality of fit for a model. It is a measure of how much better a model with more parameters fits the data. It is used to compare nested models. A nested model is a model which is a subset of another model. For example, if we have two models, the first describes the relationship between one input variable x and an output variable y and the second describes the relationship between two input variables x and z and an output variable y, then the frist model is nested within the second.

4.5 Bayesian Information Criterion (BIC)

When there is a limited number of models, the BIC is often used for model selection. Often, the model with the smallest BIC is preferred. The BIC is a penalised version of the deviance where the penalty is relevant to the number of input variables.

5 P-values

It is common to provide p-values when conducting statistical analysis. P-values are propabilities (i.e. their values always lie between zero to 1) and they show how likely certain situations are. P-values which are close to zero (usually $<= 0.05$) are more likely to occur if the study has shown something positive. It is said that the result is significant if the p-value is close to zero. On the other hand, the result is said to be non-significant if the p-value is away from zero (usually > 0.05).

6 Confidence Intervals

It is known that in statistics we use the sample data at hand to draw inferences about the entire population (i.e. all the data) and make an estimate of the value(s) we are trying to measure or predict. It is very important to present such estimate with a measure of precision. This measure of precision depends on the sample size and normally takes the form of a 95% Confidence Interval or a standard error value (the former is calculated from the latter). The 95% confidence interval gives the range of poulation parameters that the sample leads us to believe are possible. The 95% confidence interval is presented as a range of two values (a, b) and is interpreted as: we can be 95% confident that the result/effect we are trying to measure will happen by an overage of at least a and maybe as much as b

7 Examining the Variables

As we have three numerical input variables, it is appropriate to examine their statistical summaries. This is what we show in table 1.

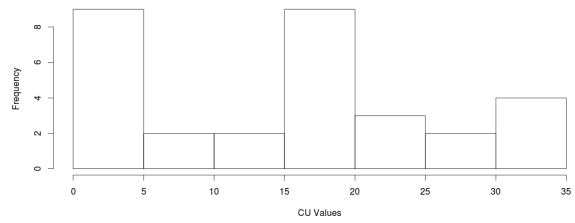
	CU	PHI	AS	Q
Valid	31	31	31	31
Missing	0	0	0	0
Mean	$17.23_{(13.24,21.21)}$	$40.13_{(38.23,42.02)}$	$21.55_{(15.53,27.58)}$	$240.3_{(166.05,314.53)}$
Median	20.00	39.00	17.00	165.0
Std. Deviation	10.87109	5.168765	16.41788	202.4052
Min	4.00	33.00	1.60	2.0
Max	35.00	49.00	65.00	820.0

Table 1: Statistical Summaries of the Variables with 95% Confidence Interval for the Mean values

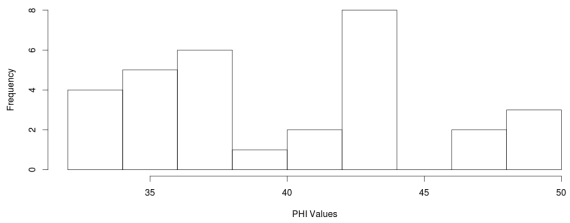
By examining the table, we observe that we have data for all the points (i.e. no missing values). Also, we notice that there is no big difference between the mean and median values of each variable, hence outliers are unlikely (outliers usually have a big influence on the difference between the mean and the median). Another values that can analyse from the table are the minimum and maximum values for each input variable. They appear to be within possible ranges for all of the three input variables. Finally, as standard deviation is an indicator of how spread out the data is, we can check the validity of our data by going 2 standard deviations on each side of the mean for the outcome variable Q. We notice that more than 95% of all values of this variable lie within that range.

8 Histograms of the Variables

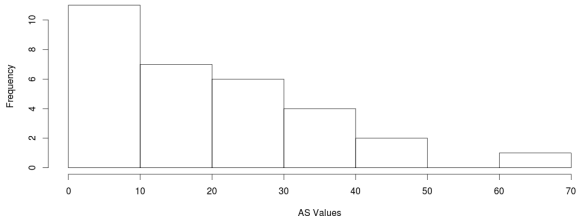
We display the histograms of the three input variables and the output variable in Figure 1. A close look at these histograms indicate some "bunching" around multiples of 5 for the CU input variable (Figure 1a) and around multiples of 10 for the AS variable (Figure 1c). Examination of the frequency distribution reveals that most values were measured to the nearest 5 XXX UNIT and 10 XXX UNIT respectively.



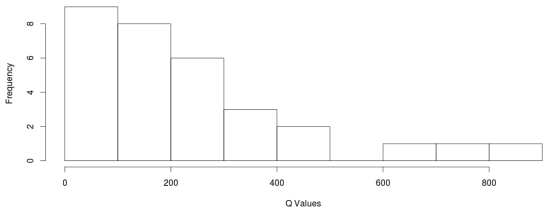
(a) Hist of CU



(b) Hist of PHI



(c) Hist of AS



(d) Hist of Q

Figure 1: Histograms of the Variables

9 Relationships between the Variables

We show a scatterplot matrix between all variables in Figure 2. The top row of this scatterplot matrix gives the scatterplots of Q against each of the other three input variables. Additionally, we show the univariate relationship between our outcome variable Q and the input variables in Table 2. From this table, we can notice that the correlation between CU and Q and between PHI and Q is significant at the $p < 0.05$ level.

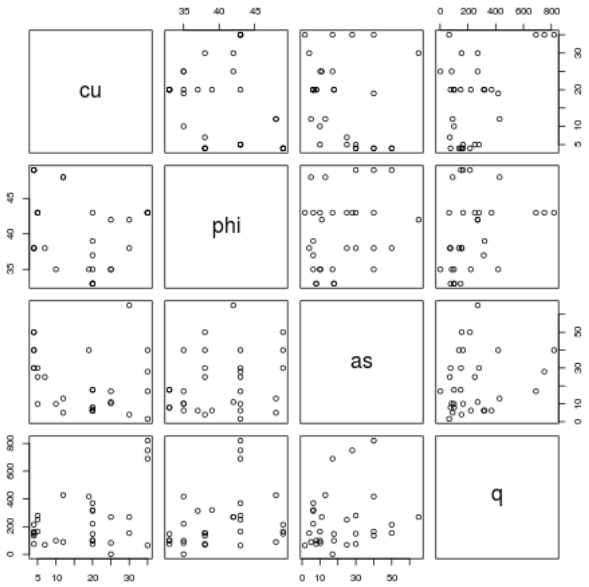


Figure 2: Scatterplot of all the variables

	Correlation (p-value) with Q	
	Pearson	Spearman
CU	0.4595897(0.009293)	0.1952728(0.2925)
AS	0.1964417(0.2895)	0.2026105(0.2743)
PHI	0.3082577(0.09158)	0.3982587(0.02649)

Table 2: Correlations between the Input Variables and Q

We also provide information about the correlations between our input variables in table 3. From this table we can observe that none of the correlations between the input variables is significant at the $p < 0.05$ level

Table 3: Correlations between Input Variables

(a) Pearson Correlation

	AS	CU	PHI
AS	1	-0.2596521(0.1584)	0.2605386(0.1569)
CU	-0.2596521(0.1584)	1	-0.2176558(0.2395)
PHI	0.2605386(0.1569)	-0.2176558(0.2395)	1

(b) Spearman Correlation

	AS	CU	PHI
AS	1	-0.3944457(0.0281)	0.204013(0.271)
CU	-0.3944457(0.0281)	1	-0.2161602(0.2428)
PHI	0.204013(0.271)	-0.2161602(0.2428)	1

10 Fitting the Models

Let us keep in mind that the ultimate aim of our final model is to determine whether there is an association between VARIABLE(s) and Q. Therefore, it would seem reasonable to model CONFOUNDER FIRST and then enter CU into the model and see whether it is associated with Q. We can do this in the opposite way by enetering CU into the model first and then follow it by PHI and AS in turn to see what impact each of them would have on the model.

10.1 Modelling the Relationship between CU and Q (Model 1)

In this section, we are going to build a simple linear regression model using just CU as input and Q as output. After using R's lm() function, the model looks as follows:

$$Q = 92.890 + 8.557 * CU \tag{1}$$

By examining equation 1, we observe that an increase, or decrease, of CU by one unit, causes an increase, or decrease, in Q by 8.557 units

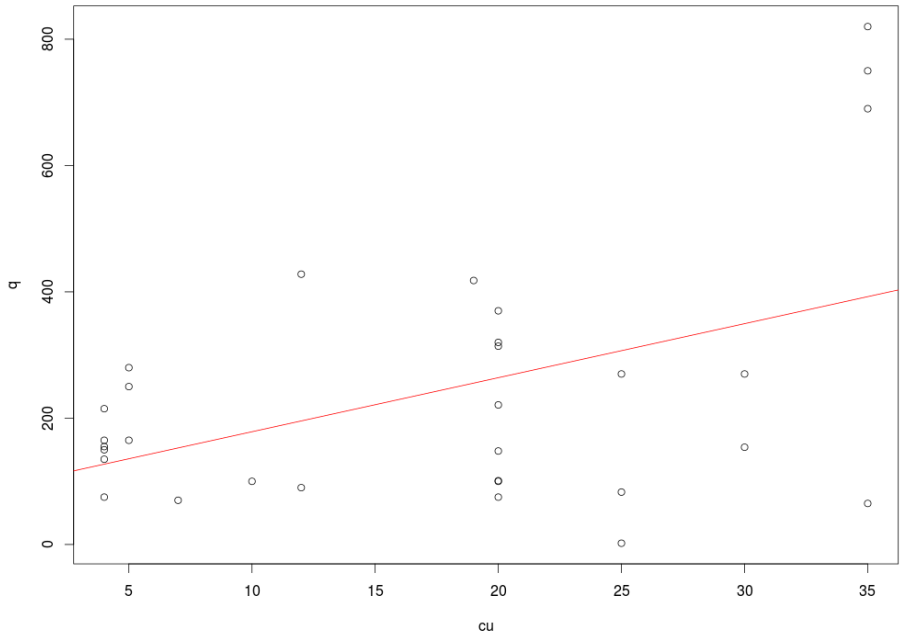


Figure 3: The Relationship between CU and Q

Now after building the model that describes the relationship between CU and Q, let us plot the input variable CU against the residuals. As Figure 4 shows, the residual values do not seem to have a particular pattern and they are randomly scattered around zero (as we stated in Section 4.3).

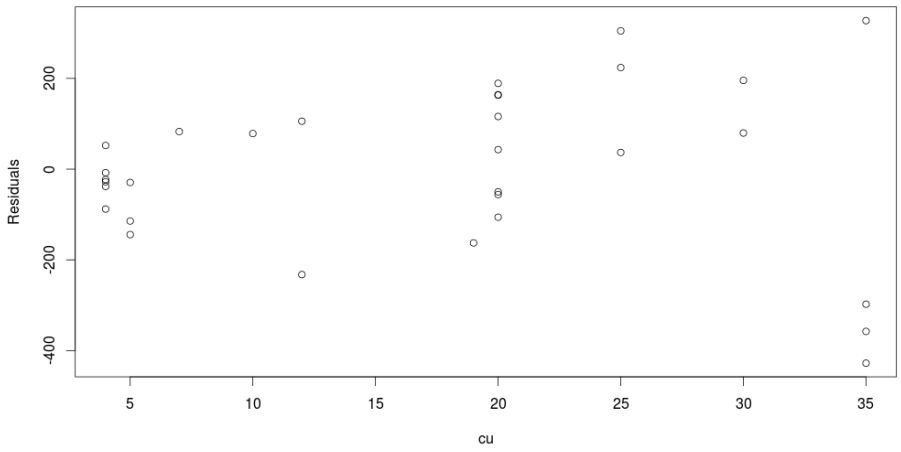


Figure 4: CU vs Residuals for Model 1

10.2 Modelling the Relationship between AS and Q (Model 2)

In this section, we are going to build a simple linear regression model using just AS as input and Q as output. After using R's lm() function, the model looks as follows:

$$Q = 188.089 + 2.422 * AS \tag{2}$$

By examining equation 2, we observe that an increase, or decrease, of AS by one unit, causes an increase, or decrease, in Q by 2.422 units

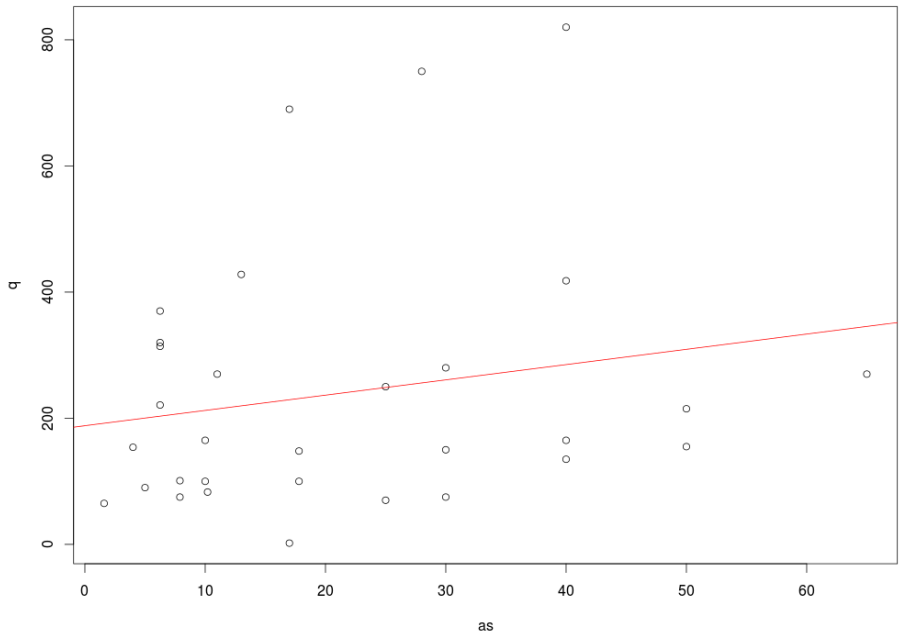


Figure 5: The Relationship between AS and Q

Now after building the model that describes the relationship between AS and Q, let us plot the input variable AS against the residuals. As Figure 6 shows, the residual values do not seem to have a particular pattern and they are randomly scattered around zero (as we stated in Section 4.3).

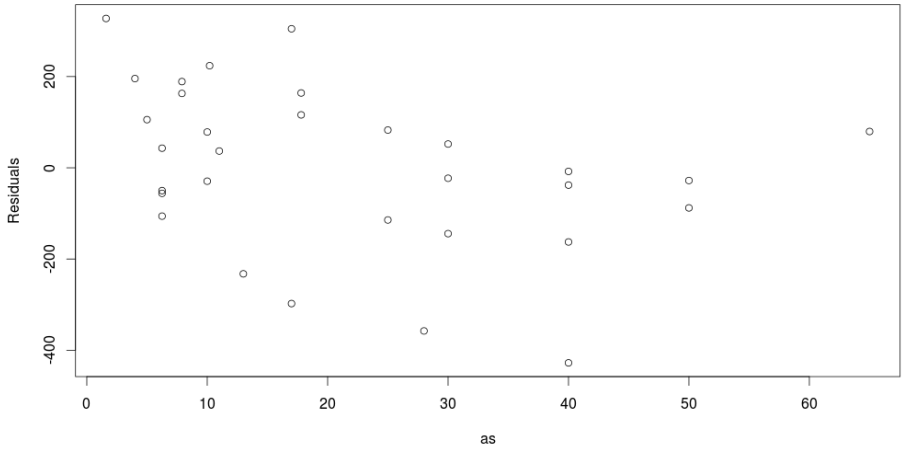


Figure 6: AS vs Residuals for Model 2

10.3 Modelling the Relationship between PHI and Q (Model 3)

In this section, we are going to build a simple linear regression model using just PHI as input and Q as output. After using R's lm() function, the model looks as follows:

$$Q = -244.11 + 12.07 * PHI \tag{3}$$

By examining equation 3, we observe that an increase, or decrease, of PHI by one unit, causes an increase, or decrease, in Q by 12.07 units

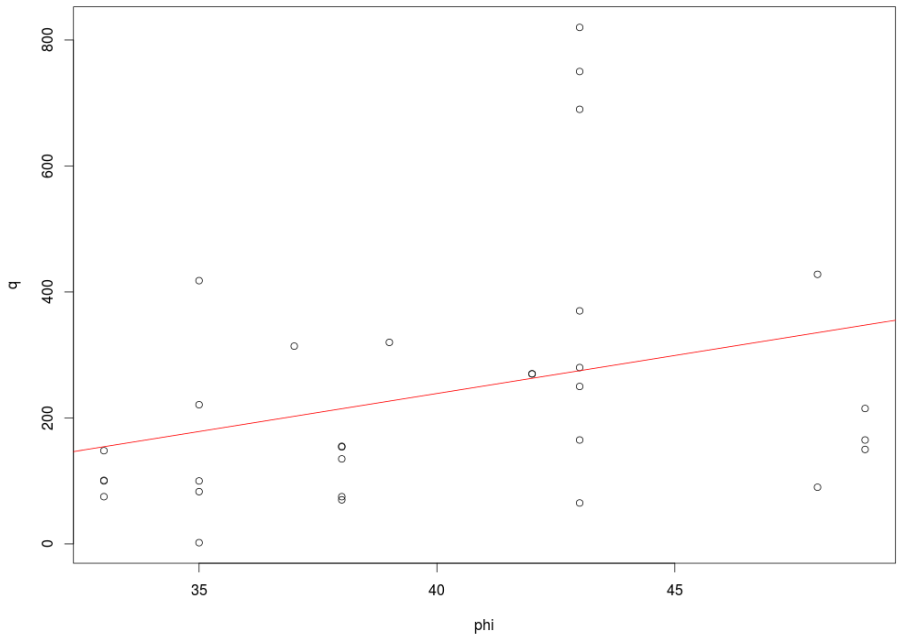


Figure 7: The Relationship between PHI and Q

Now after building the model that describes the relationship between PHI and Q, let us plot the input variable PHI against the residuals. As Figure 8 shows, the residual values do not seem to have a particular pattern and they are randomly scattered around zero (as we stated in Section 4.3).

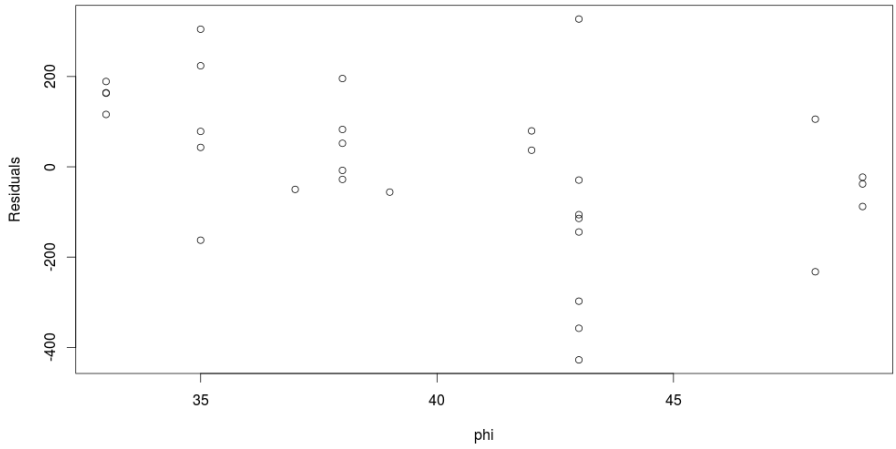


Figure 8: PHI vs Residuals for Model 3

10.4 Modelling the Relationship between CU, AS and Q (Model 4)

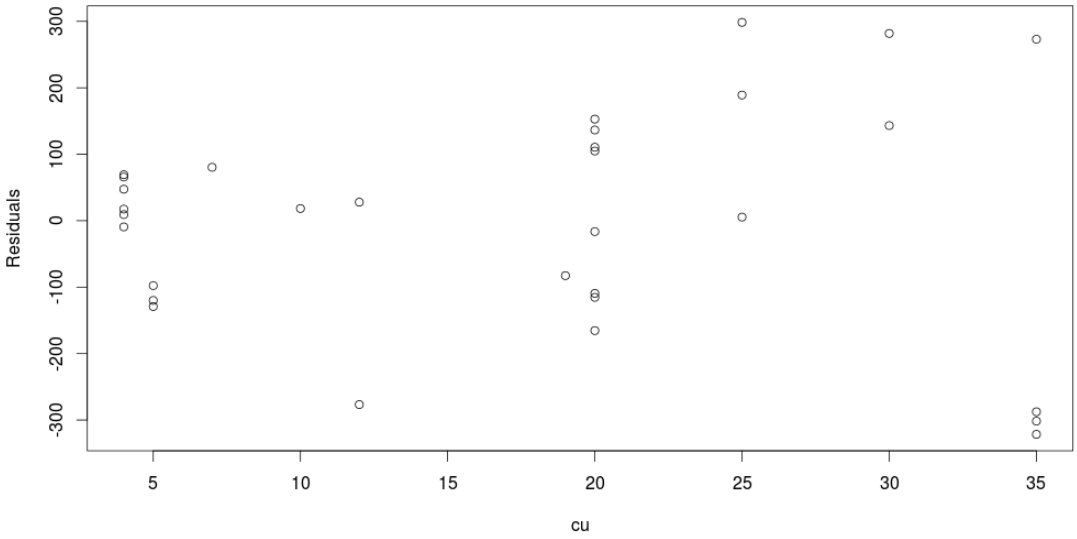
In this section, we are going to build a simple linear regression model using CU and AS as inputs and Q as output. After using R's lm() function, the model looks as follows:

$$Q = -25.286 + 10.194 * CU + 4.174 * AS \tag{4}$$

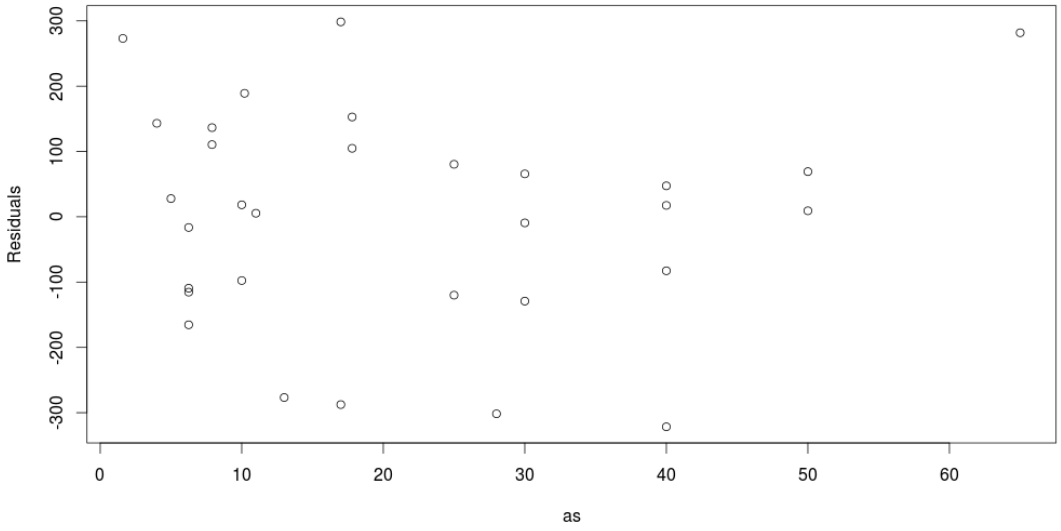
By examining equation 4, we observe that when fixing AS, an increase, or decrease, of CU by one unit, causes an increase, or decrease, in Q by 10.194 units. Similarly, when fixing CU, an increase, or decrease, of AS by one unit, causes an increase, or decrease, in Q by 4.174 units.

Now after building the model that describes the relationship between CU, AS and Q, let us plot the input variables CU and AS against the residuals. As Figure 9 shows, the residual

values do not seem to have a particular pattern and they are randomly scattered around zero (as we stated in Section 4.3).



(a) CU vs Residuals for Model 4



(b) AS vs Residuals for Model 4

Figure 9: Input Variables vs Residuals for Model 4

10.5 Modelling the Relationship between CU, PHI and Q (Model 5)

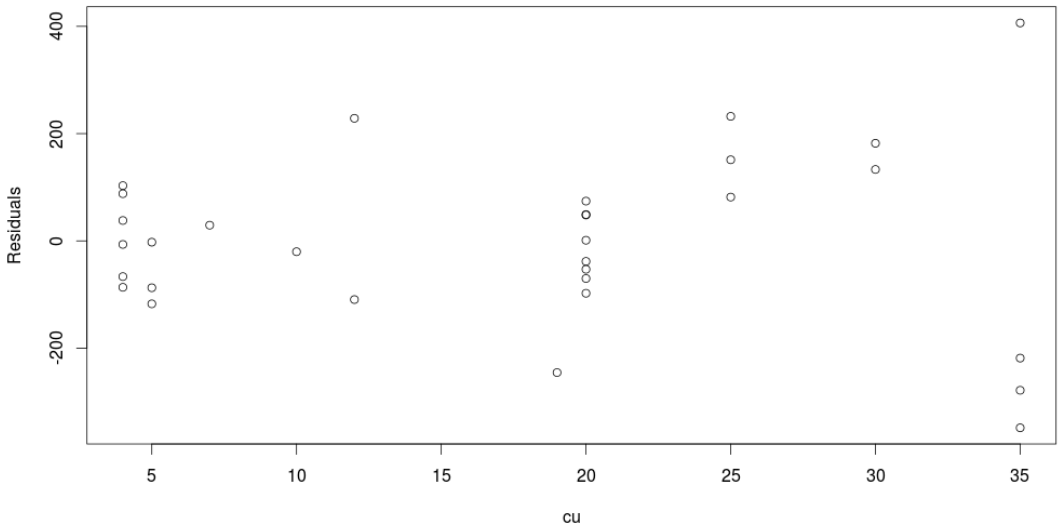
In this section, we are going to build a simple linear regression model using CU and PHI as inputs and Q as output. After using R's lm() function, the model looks as follows:

$$Q = -610.53 + 10.29 * CU + 16.78 * PHI \tag{5}$$

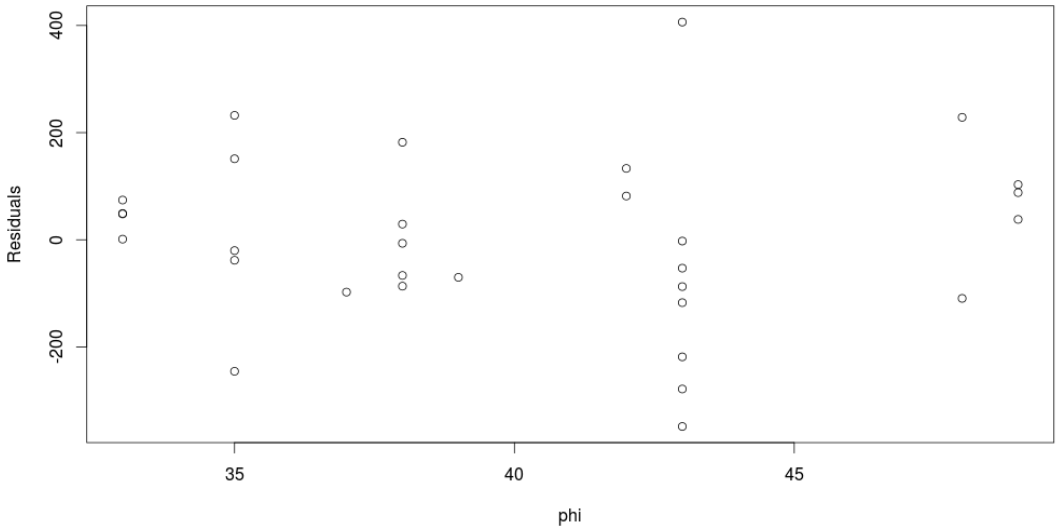
By examining equation 5, we observe that when fixing PHI, an increase, or decrease, of CU by one unit, causes an increase, or decrease, in Q by 10.29 units. Similarly, when fixing CU, an increase, or decrease, of PHI by one unit, causes an increase, or decrease, in Q by 16.78 units.

Now after building the model that describes the relationship between CU, PHI and Q, let us plot the input variables CU and PHI against the residuals. As Figure 10 shows, the residual

values do not seem to have a particular pattern and they are randomly scattered around zero (as we stated in Section 4.3).



(a) CU vs Residuals for Model 5



(b) PHI vs Residuals for Model 5

Figure 10: Input Variables vs Residuals for Model 5

10.6 Modelling the Relationship between AS, PHI and Q (Model 6)

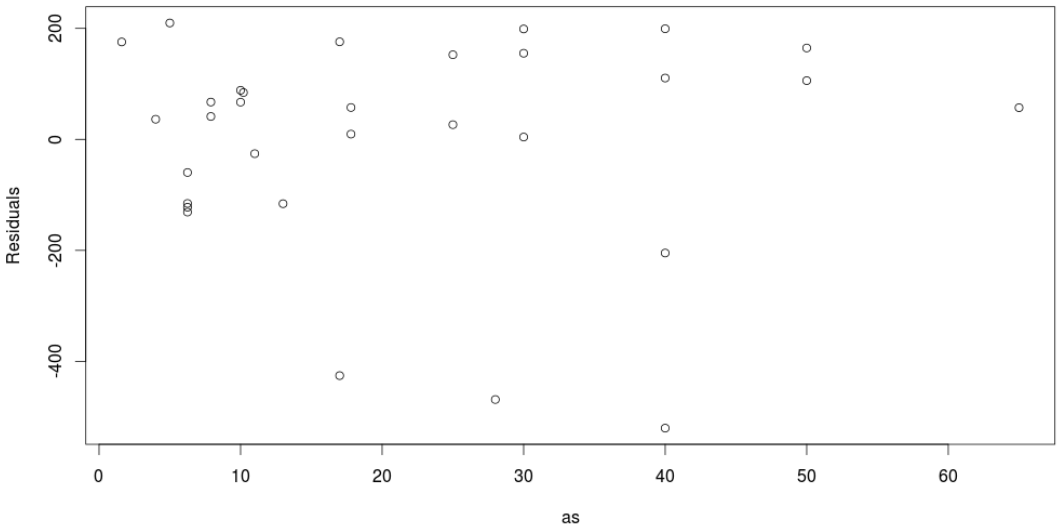
In this section, we are going to build a simple linear regression model using AS and PHI as inputs and Q as output. After using R's lm() function, the model looks as follows:

$$Q = -226.213 + 1.536 * AS + 10.8 * PHI \tag{6}$$

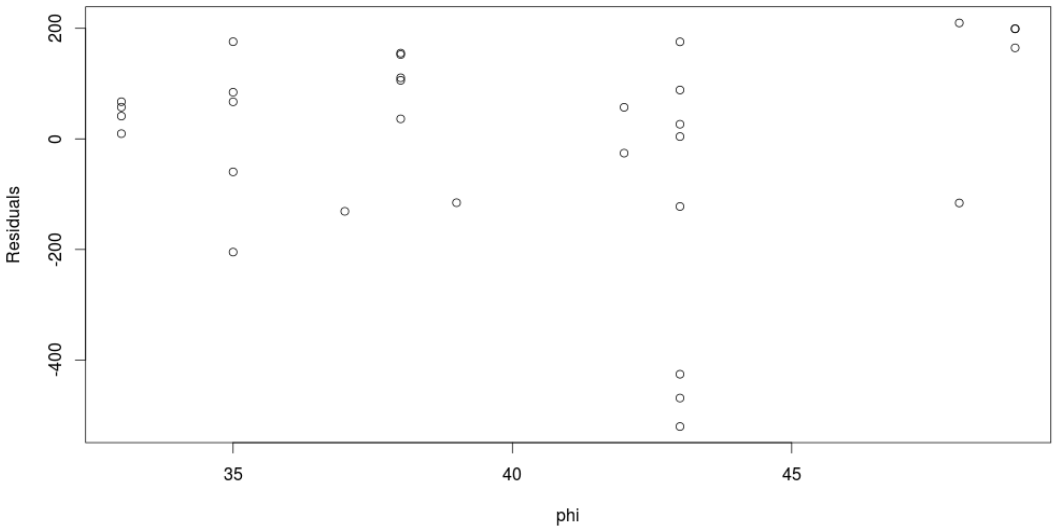
By examining equation 6, we observe that when fixing PHI, an increase, or decrease, of AS by one unit, causes an increase, or decrease, in Q by 1.536 units. Similarly, when fixing AS, an increase, or decrease, of PHI by one unit, causes an increase, or decrease, in Q by 10.8 units.

Now after building the model that describes the relationship between AS, PHI and Q, let us plot the input variables AS and PHI against the residuals. As Figure 11 shows, the residual

values do not seem to have a particular pattern and they are randomly scattered around zero (as we stated in Section 4.3).



(a) AS vs Residuals for Model 6



(b) PHI vs Residuals for Model 6

Figure 11: Input Variables vs Residuals for Model 6

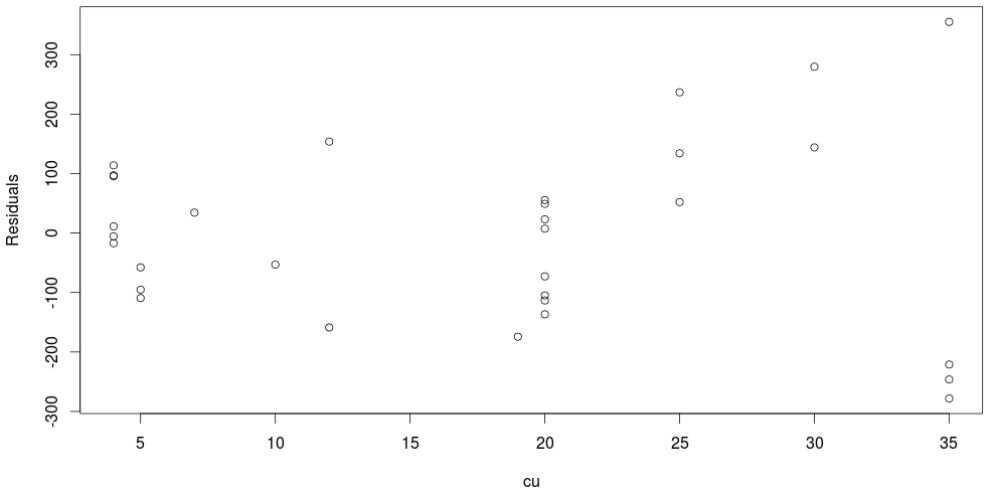
10.7 Modelling the Relationship between CU, AS, PHI and Q (Model 7)

In this section, we are going to build a simple linear regression model using CU, AS and PHI as inputs and Q as output (recall this is the purpose of this study). After using R's lm() function, the model looks as follows:

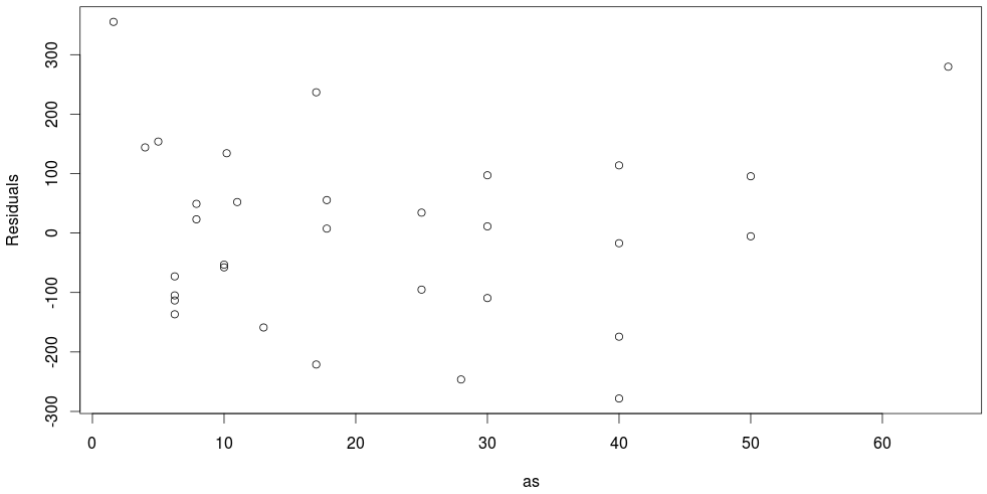
$$Q = -609.893 + 11.313 * CU + 3.167 * AS + 14.629 * PHI \tag{7}$$

By examining equation 7, we observe that when fixing PHI and AS, an increase, or decrease, of CU by one unit, causes an increase, or decrease, in Q by 11.313 units. Similarly, when fixing CU and AS, an increase, or decrease, of PHI by one unit, causes an increase, or decrease, in Q by 14.629 units. Also, when fixing CU and PHI, an increase, or decrease, of AS by one unit, causes an increase, or decrease, in Q by 3.167 units.

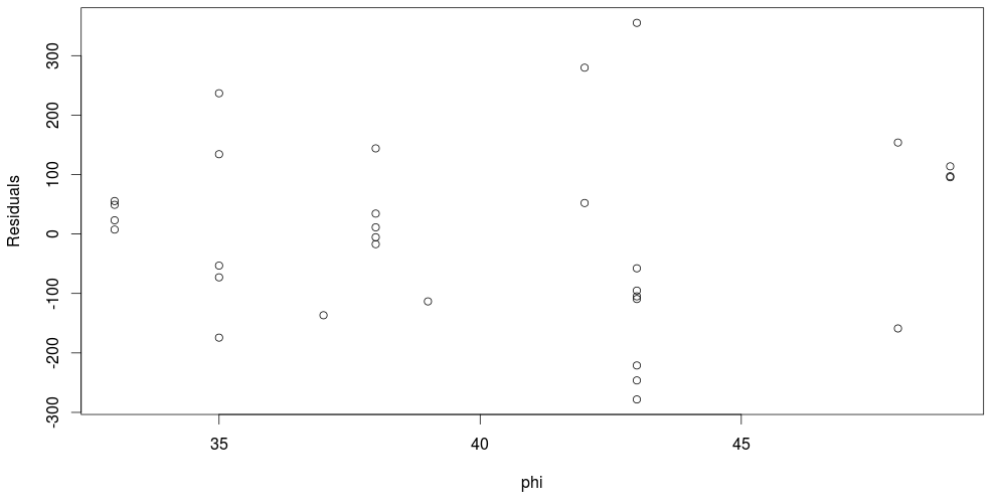
Now after building the model that describes the relationship between CU, AS, PHI and Q, let us plot the input variables CU, AS and PHI against the residuals. As Figure 12 shows, the residual values do not seem to have a particular pattern and they are randomly scattered around zero (as we stated in Section 4.3).



(a) CU vs Residuals for Model 7



(b) AS vs Residuals for Model 7



(c) PHI vs Residuals for Model 7

Figure 12: Input Variables vs Residuals for Model 7

11 Checking the Model Fit

In Section 4, we mentioned five diagnostics that can be used to examine the goodness of fit of a model. Namely, these were: Deviance, BIC, R-Squared, Adjusted R-Squared and the Residuals. We have plotted the individual input variables against the residuals when we built the models in Section 10. Table 4 shows the values of the remaining four diagnostics for the seven models.

Model	Deviance	BIC Value	R-Squared	Adj R-Squared
Model 1	969436	419.1411	0.2112227	0.1840235
Model 2	1181609	425.2766	0.03858933	0.005437238
Model 3	1112250	423.4013	0.09502283	0.06381672
Model 4	838024.3	418.0594	0.3181453	0.2694414
Model 5	754365.9	414.7992	0.3862136	0.3423717
Model 6	1094468	426.3357	0.1094908	0.04588297
Model 7	682279.8	415.1196	0.4448661	0.3831845

Table 4: Deviance, BIC, R-Squared and Adjusted R-Squared of the Seven models

As we mentioned in Section 4.5, a lower BIC indicates a better fitting model. By analysing table 4 we observe that Model 5 has the lowest BIC with Model 7 in second. Howerver, when we examine the value of R-Squared, we realise that Model 7 has the highest R-Squared amongst all the models (see Section 4.1). This gives us confidence that from amongst the seven models that we created using various combinations of the input variables, Model 7 is the best model that describes the relationship between the input variables CU, AS and PHI and the output variable Q.