

# Pricing the C's of Diamond Stones

Singfat Chu  
National University of Singapore

*Journal of Statistics Education* Volume 9, Number 2 (2001)

Copyright © 2001 by Singfat Chu, all rights reserved.

This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the author and advance notification of the editor.

---

**Key Words:** Categorical variables; Data transformation; Multiple linear regression; Standardized residuals.

## Abstract

Many statistical problems can be satisfactorily resolved within the framework of linear regression. Business students, for example, employ linear regression to uncover interesting insights in the fields of Finance, Marketing, and Human Resources, among others. The purpose of this paper is to demonstrate how several concepts arising in a typical discussion of multiple linear regression can be motivated through the development of a pricing model for diamond stones. Specifically, we use data pertaining to 308 stones listed in an advertisement to construct a model, which educates us on the relative pricing of caratage and the different grades of clarity and colour.

## 1. Introduction

Regression analysis is a most versatile tool in our students' statistical arsenal. It is perhaps the most useful statistical technique employed by them during their academic experience and later in their professional endeavours. Having gone through the complexities of independent samples t-test and ANOVA, many students are relieved when they realise that the comparison of group means can actually be conducted within the unified framework of the regression model. The latter also offers flexibility and transparency in handling exogenous factors.

In March 2000, I tasked my MBA students to develop a sensible pricing model for diamond stones using data that appeared in an advertisement in Singapore's *Business Times* edition of February 18, 2000. An example of such an advertisement appears in [Figure 1](#). The analysis was to focus on data pertaining to  $n = 308$  Round diamond stones (the other less popular shapes being Heart, Pear, Princess, Marquise, Emerald). More recently, I have redesigned the application as an in-class case study supported by live usage of the Microsoft Excel® software. The allure of the application has generated much enthusiasm and discussion among the students. They have also learned that the resolution of a satisfactory statistical pricing model is achieved after a multi-round investigation process.

The Millennium SPARK  
YEAR 2000  
International Certificates

**GIA**  
Gem Trade Laboratory, Inc.  
New York, USA  
A wholly owned subsidiary of the Gemological Institute of America

**HRD**  
Hoge Raad voor Diamant  
Antwerp, Belgium  
Certification and supervision  
of the University of Antwerp

**IGI**  
International Gemological  
Institute, Antwerp, Belgium  
Reports issued by IGI, Antwerp  
Belgium. Offered for sale as

Heart Pear Princess Round Marquise Emerald

All prices are inclusive of free pendant, ring and ear studs settings (set & collected on the spot).  
G.S.T. absorbed by Establishment

Figure 1

Figure 1. An Advertisement for Diamonds.

## 2. What Price Diamond Stones?

The website [www.adiamondisforever.com](http://www.adiamondisforever.com) educates the layperson on the factors that influence the price of a diamond stone. These are the 4 C's: Carat, Clarity, Colour and Cut.

The weight of a diamond stone is indicated in terms of carat units. One carat is equivalent to 0.2 grams. All other things being equal, larger diamond stones command higher prices in view of their rarity.

Being products of Nature, diamonds have "birthmarks" or inclusions only visible under a jeweller's magnifying glass or a microscope. Diamonds with no inclusion under a loupe with a 10 power magnification are labelled IF ("internally flawless"). Lesser diamonds are categorised in descending order as "very very slightly imperfect" VVS1 or VVS2 and "very slightly imperfect" VS1 or VS2.

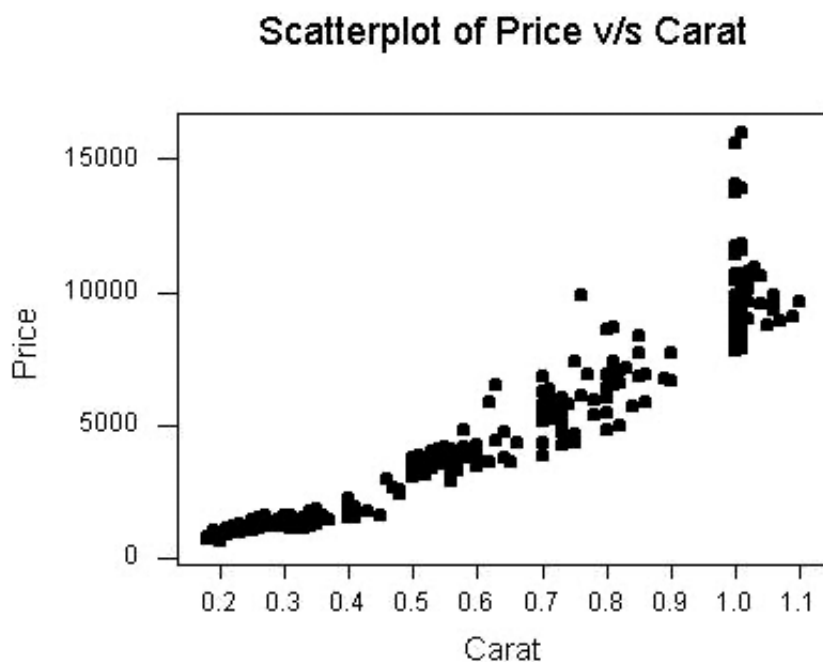
The most prized diamonds display colour purity. They are not contaminated with yellow or brown tones. Top colour purity attracts a grade of D. Subsequent degrees of colour purity are rated E, F, G, ... all the way down the alphabet ladder.

The cut (or faceting) of a raw diamond stone relies on the experience and the craftsmanship of the diamond cutter. The optimal cut should neither be too deep nor too shallow for it will impede the trajectory of light and thereby the brilliance or “fire” of a diamond stone.

To assist shoppers, independent certification bodies assay diamond stones and provide each of them with a certificate listing their caratage and their grades of clarity, colour and cut. The newspaper advertisement however only provided, for each stone, details on the certification body and its assessment of the caratage, clarity and colour of the stones. Three certification bodies were mentioned in the advertisement, namely New York based Gemmological Institute of America (GIA) and Antwerp based International Gemmological Institute (IGI) and Hoge Raad Voor Diamant (HRD). Their reputations could be a factor in the pricing of the diamond stones.

### 3. First Attack on the Pricing Problem

Given the information in the dataset, a multiple linear regression (MLR) model is a natural path to explore. Generally speaking, one would expect the price (denoted in Singapore dollars) of a stone to move in tandem with the caratage. However, the relationship may not be linear as heavier stones are more prized than the lighter ones. An examination of the scatter plot of Price against Carats would therefore be enlightening.

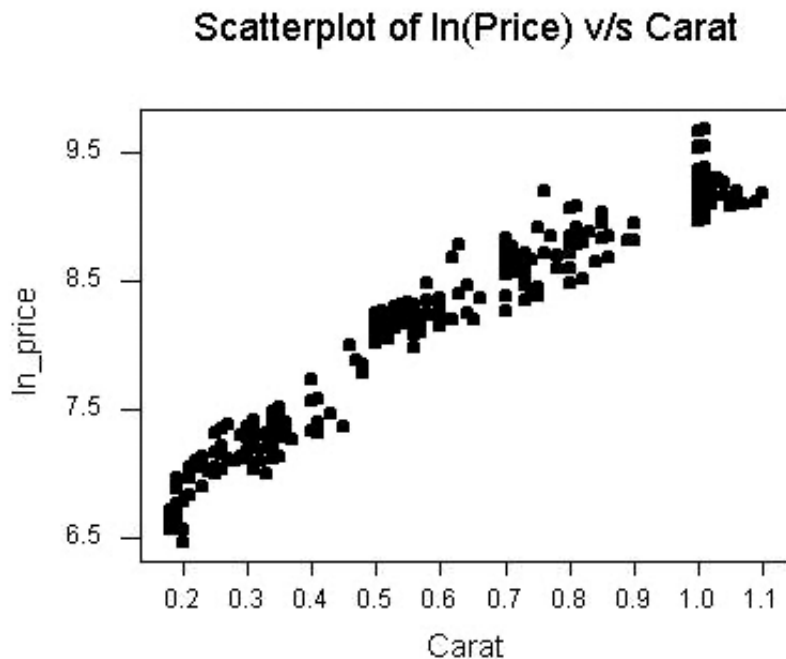


[Figure 2](#)

Figure 2. Price Against Carat.

Clearly, there is a relationship but the trend appears to fan out. This indicates higher price volatility for the heavier stones, especially those above 1 carat. Unless we transform the data, we would most likely not satisfy the homoscedasticity assumption of linear regression. A transformation that is recommended in

similar situations is the logarithm of prices. This is illustrated below.



[Figure 3](#)

Figure 3. Ln(Price) Against Carat.

The relationship between Carat and the logarithms of Price appears more homoscedastic compared to the first scatter plot. This suggests that it would be more judicious to employ  $\ln(\text{Price})$  in lieu of Price in developing a linear regression model.

Next we have to insert clarity, colour and the identity of the certification body in the regression model. Students should notice that these are all categorical in nature. Therefore the operational hurdle facing them is the following:

*Discussion 1: How should the ordinal data be coded?*

In the case of ordinal data like clarity (ditto for colour), some students may be tempted to employ, for example, VS2=1, VS1=2, VVS2=3, VVS1=4 and IF=5. A discussion would therefore have to be engaged on why this is not suitable.

The MINITAB® output and accompanying residual plots from the first attack on the data are reproduced below. Selecting clarity grade VS2 as my baseline category, I coded four indicator variables to help me infer on the difference between VS2 and each of VS1, VVS2, VVS1 and IF. Likewise, I defined colour I as the baseline and compared it to the other five colours using five indicator variables. Instructors may use Discussion 2 to guide their classes in assessing the results.

*Discussion 2: Is the regression model useful? This requires students to assess whether (a) the model has predictive power, (b) the estimates of the regression slopes are sensible, especially for the ordinal data, and (c) the standard assumptions of MLR are met. Students can also be queried on the advantage of*

*scrutinizing standardized as opposed to raw residuals.*

The regression equation is

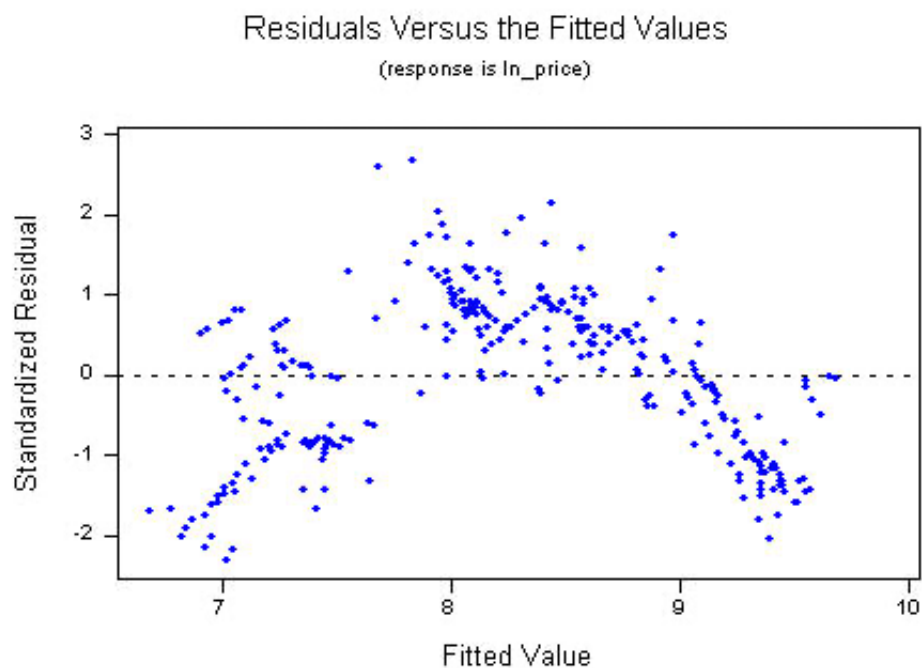
$$\ln\_price = 6.08 + 2.86 \text{ Carat} + 0.417 \text{ D} + 0.387 \text{ E} + 0.310 \text{ F} + 0.210 \text{ G} + 0.129 \text{ H} \\ + 0.299 \text{ IF} + 0.298 \text{ VVS1} + 0.202 \text{ VVS2} + 0.0966 \text{ VS1} + 0.0089 \text{ GIA} \\ - 0.174 \text{ IGI}$$

Predictor	Coef	StDev	T	P
Constant	6.07724	0.04809	126.37	0.000
Carat	2.85501	0.03697	77.23	0.000
D	0.41656	0.04138	10.07	0.000
E	0.38705	0.03082	12.56	0.000
F	0.31020	0.02748	11.29	0.000
G	0.21021	0.02836	7.41	0.000
H	0.12868	0.02852	4.51	0.000
IF	0.29854	0.03330	8.96	0.000
VVS1	0.29783	0.02810	10.60	0.000
VVS2	0.20192	0.02534	7.97	0.000
VS1	0.09661	0.02492	3.88	0.000
GIA	0.00886	0.02086	0.42	0.672
IGI	-0.17385	0.02867	-6.06	0.000

S = 0.1382      R-Sq = 97.2%      R-Sq(adj) = 97.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	12	197.939	16.495	863.64	0.000
Residual Error	295	5.634	0.019		
Total	307	203.574			



[Figure 4](#)

Figure 4. Residual Plot.

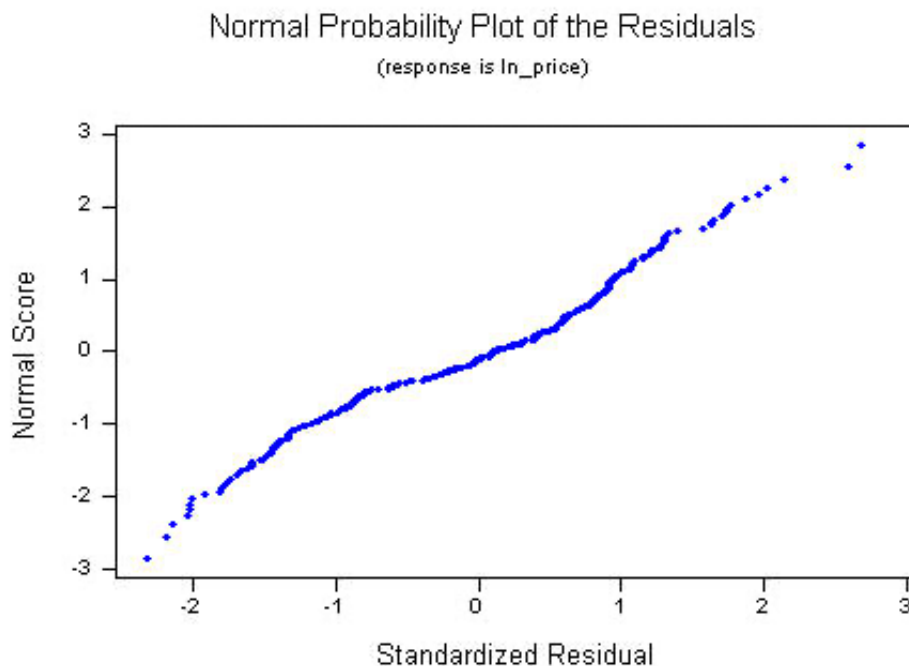


Figure 5

Figure 5. Normal Plot.

The students' verdict should be that although the model has predictive power and the slopes adhere to the hierarchy of the grades of colour and clarity, the dome-like scatter in the residual plot is a cause for concern. The normality assumption, however, appears to be less problematic.

*Discussion 3: What remedial action(s) can be undertaken?*

## 4. Remedial Actions

The residual plot indicates that the regression model underestimates prices at both ends of the price range and overestimates the midrange prices.

This insight opens up several vistas for exploration. One possibility is to segregate the stones according to caratage. For instance, Figure 2 suggests that the stones may be divided into 3 clusters, say less than 0.5 carats ("small"), 0.5 to less than 1 carat ("medium") and 1 carat and over ("large"). Separate regression models may be constructed for each cluster. The disadvantage of this approach is that results may not be consistent across the 3 clusters as these do not have an even spread of the grades of colour and clarity. This leads to the following poser,

*Discussion 4: Can we construct a unified regression model that will cover all the 308 stones and will possibly deliver different pricing structures for the 3 clusters just defined?*

This is where students would reckon that indicator variables coding the above three caratage ranges and

their interactions with carats (to reflect different slopes) will have to be employed. This avenue has been explored in my classes. Here is the MINITAB® output where “small” was defined as the baseline caratage cluster and where the coefficient for med\*carat (ditto for large\*carat) is the average difference in incremental price per carat unit between “small” and “medium” stones.

The regression equation is

$$\begin{aligned} \ln\_price = & 5.53 + 4.26 \text{ Carat} + 0.434 \text{ D} + 0.349 \text{ E} + 0.273 \text{ F} + 0.188 \text{ G} + 0.108 \text{ H} \\ & + 0.311 \text{ IF} + 0.213 \text{ VVS1} + 0.134 \text{ VVS2} + 0.0682 \text{ VS1} + 0.00770 \text{ GIA} \\ & - 0.0167 \text{ IGI} + 0.946 \text{ med} + 2.38 \text{ large} - 1.77 \text{ med*carat} \\ & - 3.26 \text{ large*carat} \end{aligned}$$

Predictor	Coef	StDev	T	P
Constant	5.5307	0.03288	168.22	0.000
Carat	4.2572	0.08550	49.79	0.000
D	0.4336	0.01690	25.66	0.000
E	0.3487	0.01255	27.78	0.000
F	0.2728	0.01114	24.49	0.000
G	0.1879	0.01152	16.31	0.000
H	0.1079	0.01148	9.39	0.000
IF	0.3114	0.01354	22.99	0.000
VVS1	0.2133	0.01154	18.49	0.000
VVS2	0.1342	0.01035	12.96	0.000
VS1	0.0682	0.01006	6.78	0.000
GIA	0.00770	0.008473	0.91	0.364
IGI	-0.0167	0.01218	-1.37	0.171
med	0.9460	0.03909	24.20	0.000
large	2.3760	0.3198	7.43	0.000
med*carat	-1.7655	0.09350	-18.88	0.000
large*carat	-3.2600	0.3234	-10.08	0.000

S = 0.05540      R-Sq = 99.6%      R-Sq(adj) = 99.5%

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	16	202.680	12.668	4126.79	0.000
Residual Error	291	0.893	0.003		
Total	307	203.574			

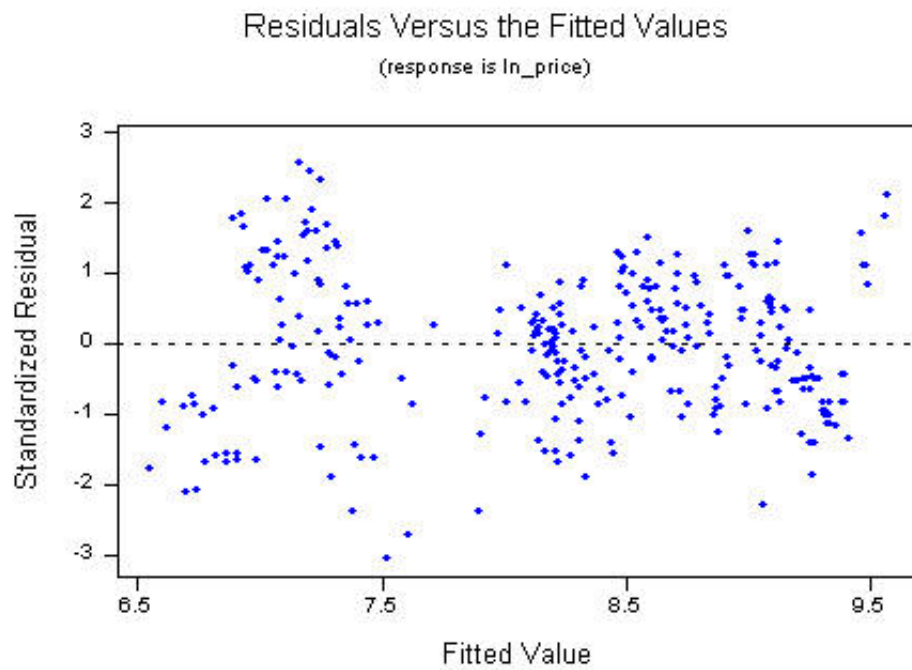
[Figure 6](#)

Figure 6. Residual Plot.

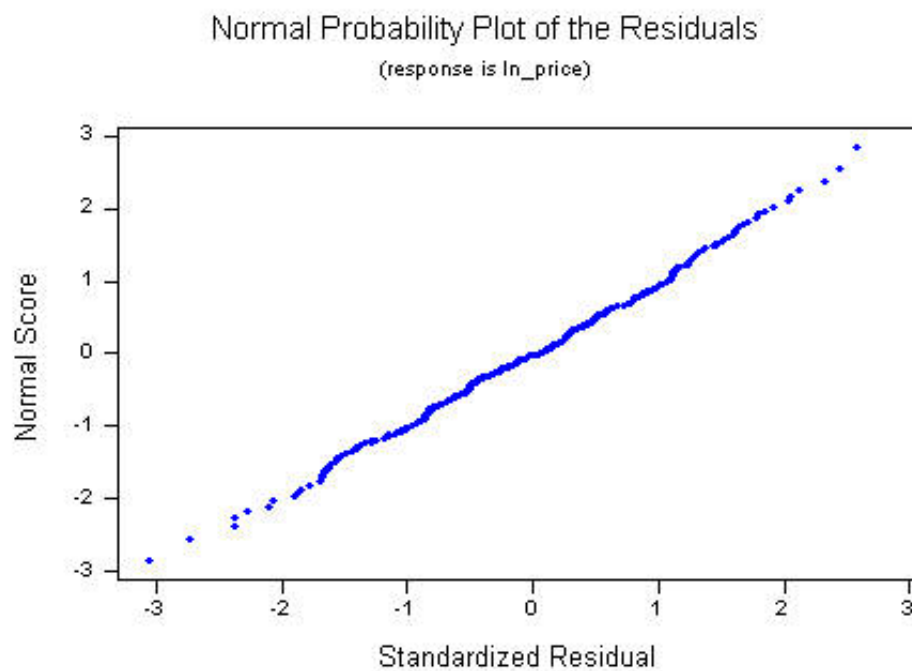
[Figure 7](#)

Figure 7. Normal Plot.

*Discussion 5: Is this regression model satisfactory? Are the standard assumptions of linear regression*



*validated? Are the numerical estimates sensible? Interpret the interaction parameter med\*carat. Which is more highly valued: colour or clarity? What can we infer on the incremental pricing of caratage in the 3 clusters? All other things being equal, what is the average price difference between a grade D diamond and another one graded (a) I (b) E? etc. All other things being equal, are there price differences amongst the stones appraised by the GIA, IGI and HRD?*

Another remedial option, which avoids the subjectivity of cluster definitions, is to employ the square of carat, as suggested by the curvature in Figure 3. The statistical output and diagnostic plots are shown below:

The regression equation is

$$\ln\_price = 5.31 + 5.67 \text{ Carat} + 0.443 \text{ D} + 0.363 \text{ E} + 0.287 \text{ F} + 0.198 \text{ G} + 0.104 \text{ H} \\ + 0.177 \text{ IF} + 0.226 \text{ VVS1} + 0.143 \text{ VVS2} + 0.0757 \text{ VS1} + 0.00622 \text{ GIA} \\ - 0.0192 \text{ IGI} - 2.10 \text{ Caratsq}$$

Predictor	Coef	StDev	T	P
Constant	5.30634	0.02961	179.20	0.000
Carat	5.67062	0.07928	71.52	0.000
D	0.44261	0.01774	24.95	0.000
E	0.36336	0.01322	27.48	0.000
F	0.28662	0.01179	24.31	0.000
G	0.19757	0.01215	16.26	0.000
H	0.10351	0.01224	8.46	0.000
IF	0.17670	0.01259	14.03	0.000
VVS1	0.22617	0.01220	18.54	0.000
VVS2	0.14348	0.01098	13.07	0.000
VS1	0.07571	0.01069	7.08	0.000
GIA	0.006223	0.008938	0.70	0.487
IGI	-0.01919	0.01300	-1.48	0.141
Caratsq	-2.10292	0.05802	-36.24	0.000

S = 0.05920      R-Sq = 99.5%      R-Sq(adj) = 99.5%

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	13	202.543	15.580	4445.36	0.000
Residual Error	294	1.030	0.004		
Total	307	203.574			

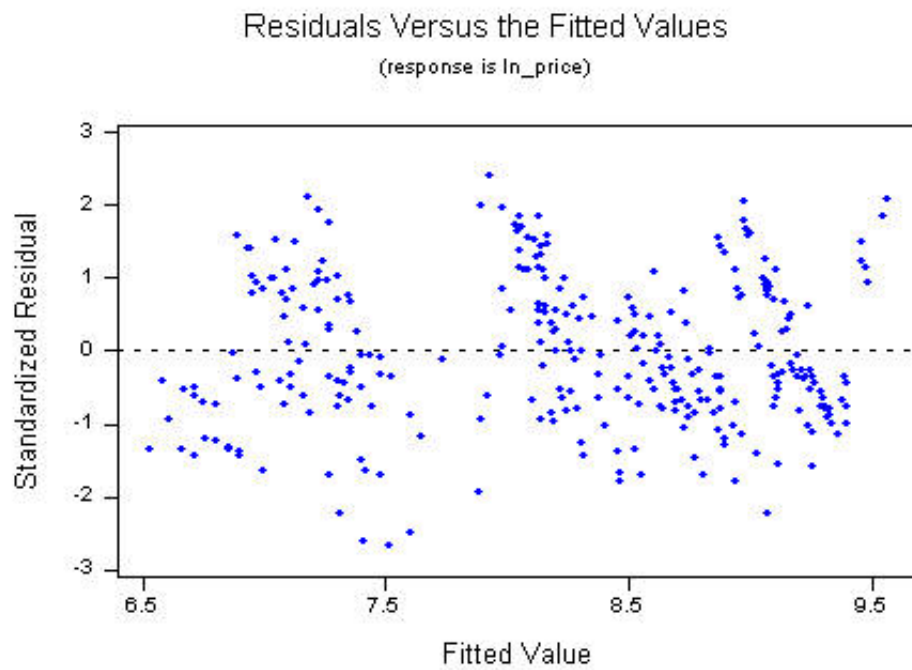
[Figure 8](#)

Figure 8. Residual Plot.

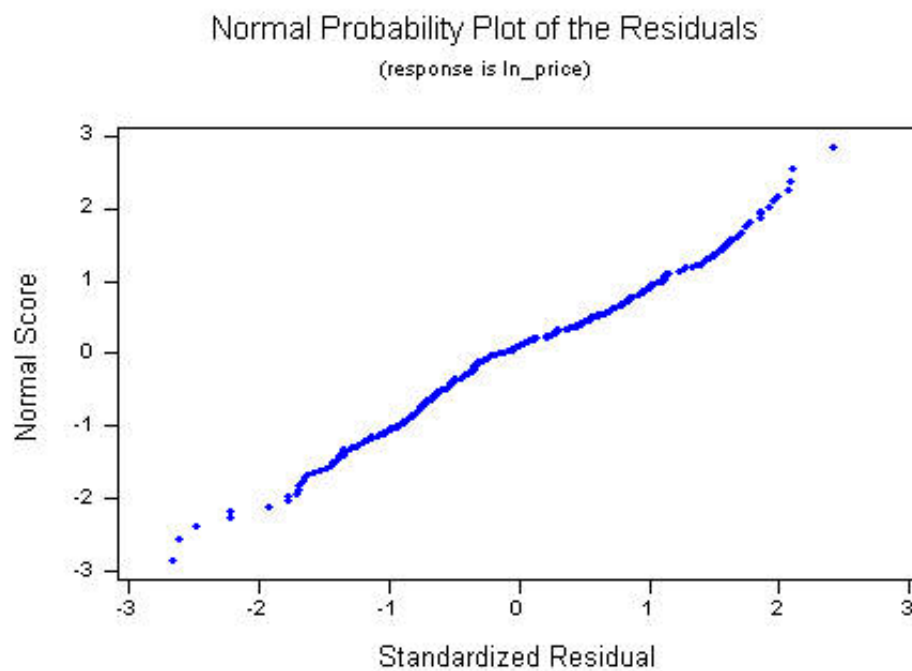
[Figure 9](#)

Figure 9. Normal Plot.

*Discussion 6: Which remedial option is preferable? Students here would scrutinize the adjusted R-*

*squares, the standard deviation of the residuals, the residual plots and the sensibilities of the regression estimates. The issue of interpretability may also be raised. Specifically, do we learn more about pricing using the variables medium, large, med\*carat and large\*carat as opposed to caratsq?*

## 5. Conclusion

In many textbook exercises, students are provided with neat datasets where often “everything works out” at first attempt. In real life, this is rarely the case. Students should be exposed to real-life datasets where they would have to exercise judgment before arriving at practical results.

In this regression application, students get to infer the pricing of the caratage and the grades of the colour and clarity of diamond stones. Unlike the hard sciences where physical laws exist to guide knowledge, statistics is about the only tool that students in business or the social sciences can use to get a grip on phenomena arising in their disciplines.

Instructors only interested in a simple linear regression application linking caratage to price may refer to an earlier publication ([jse.amstat.org/v4n3/datasets.chu.html](https://jse.amstat.org/v4n3/datasets.chu.html))

## 6. Getting the Data

The basic data are collated in the file [4Cdata.txt](#). The dataset with the indicator or "dummy" codes and transformed variables, as employed in the above analyses, is in [4C1data.txt](#). A synopsis of the application and a description of the variables are provided in the [4C.txt](#) file.

---

### Appendix to Variables in 4C.dat.txt

```
Columns
1 - 4   Carat - Weight of diamond stones in carat units
6       Colour - D, E, F, G, H or I
8 - 11  Clarity - IF, VVS1, VVS2, VS1 or VS2
13 - 15 Certification Body - GIA, IGI or HRD
18 - 21 Price (Singapore $)
```

### Appendix to Variables in 4C1.dat.txt

```
Columns
1 - 4   Carat - Weight of diamond stones in carat units
6       Indicator for colour D
8       Indicator for colour E
10      Indicator for colour F
12      Indicator for colour G
14      Indicator for colour H
16      Indicator for clarity IF
18      Indicator for clarity VVS1
20      Indicator for clarity VVS2
22      Indicator for clarity VS1
24      Indicator for certification body GIA
26      Indicator for certification body IGI
28      Indicator for medium stones between 0.5 to less than 1 carat
```

30	Indicator for large stones weighing 1 carat or more
32 - 35	Interaction variable med*carat
37 - 40	Interaction variable large*carat
42 - 48	Carat squared
50 - 53	Price (Singapore \$)
55 - 65	Ln(Price)

---

Singfat Chu  
Faculty of Business Administration  
National University of Singapore  
10 Kent Ridge Crescent  
Singapore 119260

[fbachucl@nus.edu.sg](mailto:fbachucl@nus.edu.sg)

---

[Volume 9 \(2001\)](#) | [Archive](#) | [Index](#) | [Data Archive](#) | [Information Service](#) | [Editorial Board](#) | [Guidelines for Authors](#) | [Guidelines for Data Contributors](#) | [Home Page](#) | [Contact JSE](#) | [ASA Publications](#)