Diagnostics for hedonic models

using an example for cars

(Hedonic Regression)

Kevin McCormack

September 2003

Contents **Page Number**

1. Summarising relationships

1.1 Introduction

Statistical analysis is used to document relationships among variables. Relationships that yield dependable predications can be exploited commercially or used to eliminate waste from processes. A marketing study done to learn the impact of price changes on coffee purchases is a commercial use. A study to document the relationship between moisture content of raw material and yield of usable final product in a manufacturing plant can result from finding acceptable limits on moisture content and working with suppliers to provide raw material reliably within these litmus. Such efforts can improve the efficiency of manufacturing process.

We strive to formulate statistical problems in terms of comparisons. For example, the marketing study in the preceding paragraph was conducted by measuring coffee purchases when prices were set a several different levels over a period of time. Similarly, the raw material study was conducted by comparing the yields from batches of raw materials that exhibited moisture content.

1.2 Scatterplots

Scatterplots display statistical relationship relationships between two metric variables (e.g. price and cc) in this section the details of scatterplotting are presented using the data in *Table 1*. The data were collected for used in compiling the *New Car Index* in the Irish CPI

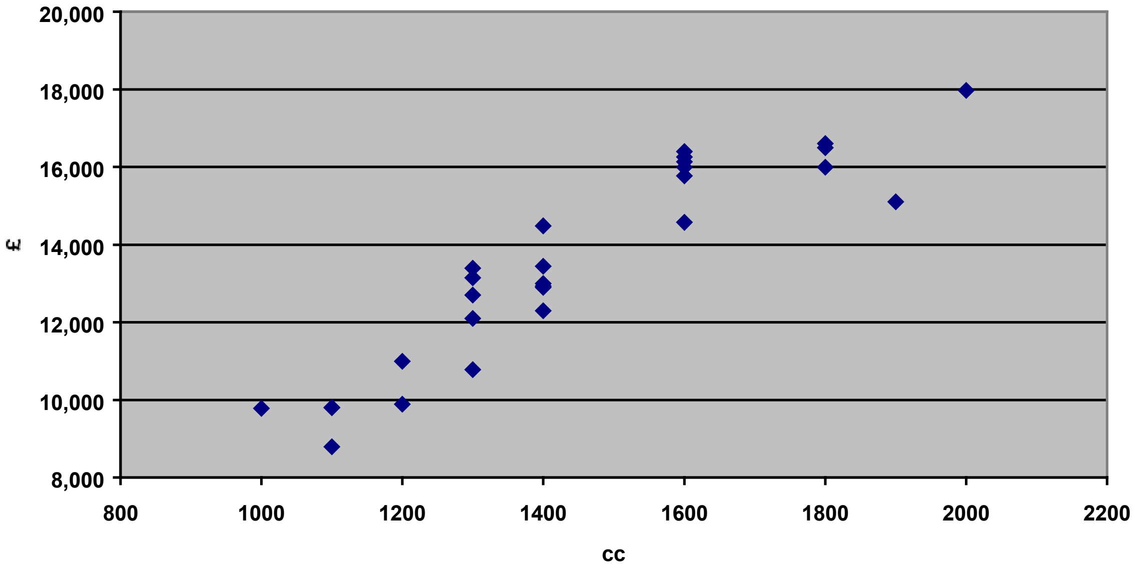
Table 1. Characteristics of cars used in the Irish New Car Price Index

	Price	CC	No. of doors	Horse Power	Weight	Length	Power steering	ABS	Air bags
	£				Kg	cm			
	(p)	(cc)	(d)	(ps)	(w)	(l)	(pst)	(abs)	(ab)
1. Toyota Corolla 1.3L Xli Saloon	13,390	1300	4	78	1200	400	1	0	0
2. Toyota Carina 1.6L SLi Saloon	15,990	1600	4	100	1400	450	1	1	1
3. Toyota Starlet 1.3L	10,780	1300	3	78	1000	370	0	0	0
4. Ford Fiesta Classic 1.3L	9,810	1100	3	60	1000	370	0	0	1
5. Ford Mondeo LX 1.6I	15,770	1600	4	90	1400	450	1	1	1
6. Ford Escort CL 1.3I	12,095	1300	5	75	1200	400	1	0	0
7. Mondeo CLX 1.6i	16,255	1600	5	90	1400	450	1	1	1
8. Opel Astra GL X1.4NZ	12,935	1400	5	60	1200	400	1	0	0
9. Opel Corsa City X1.2SZ	9,885	1200	3	45	1000	370	1	0	1
10. Opel Vectra GL X1.6XEL	16,130	1600	4	100	1400	450	1	1	1
11. Nissan Micra 1.0L	9,780	1000	3	54	1000	370	0	0	0
12. Nissan Almera 1.4GX 5sp	13,445	1400	5	87	1200	400	1	0	0
13. Nissan Primera SLX	16,400	1600	4	100	1400	450	1	1	1
14. Fiat Punto 55 SX	8,790	1100	3	60	1000	370	0	0	0
15. VW Golf CL 1.4	12,995	1400	5	60	1200	400	1	0	0
16. VW Vento CL 1.9D	15,100	1900	4	64	1400	450	1	1	1
17. Mazda 323 LX 1.3	12,700	1300	3	75	1200	400	1	0	0
18. Mazda 626 GLX 2.0I S/R	17,970	2000	5	115	1400	450	1	1	1
19. Mitsubishi Lancer 1.3 GLX	13,150	1300	4	74	1200	400	1	0	1
20. Mitsubishi Gallant 1.8 GLSi	16,600	1800	5	115	1400	450	1	1	1
21. Peugeot 106 XN 1.1 5sp	9,795	1100	5	45	1000	370	0	0	0
22. Peugeot 306 XN 1.4 DAB	12,295	1400	4	75	1200	400	1	0	0
23. 406 SL 1.8 DAB S/R	16,495	1800	4	112	1400	450	1	1	1
24. Rover 214 Si	12,895	1400	3	103	1200	400	1	0	1
25. Renault Clio 1.2 RN	10,990	1200	5	60	1000	370	1	0	1
26. Renault Laguna	15,990	1800	5	95	1400	450	1	1	1
27. Volvo 440 1.6 Intro Version	14,575	1600	5	100	1400	450	1	0	1
28. Honda Civic 1.4I SRS	14,485	1400	4	90	1200	400	1	0	0

Scatterplots are used to try to discover a tendency for plotted variables to be related in a simple way. Thus the more the scatterplot reminds us of a mathematical curve, the more closely related we infer the variables are. In the scatterplot a *direct* relationship between the two variables is inferred.

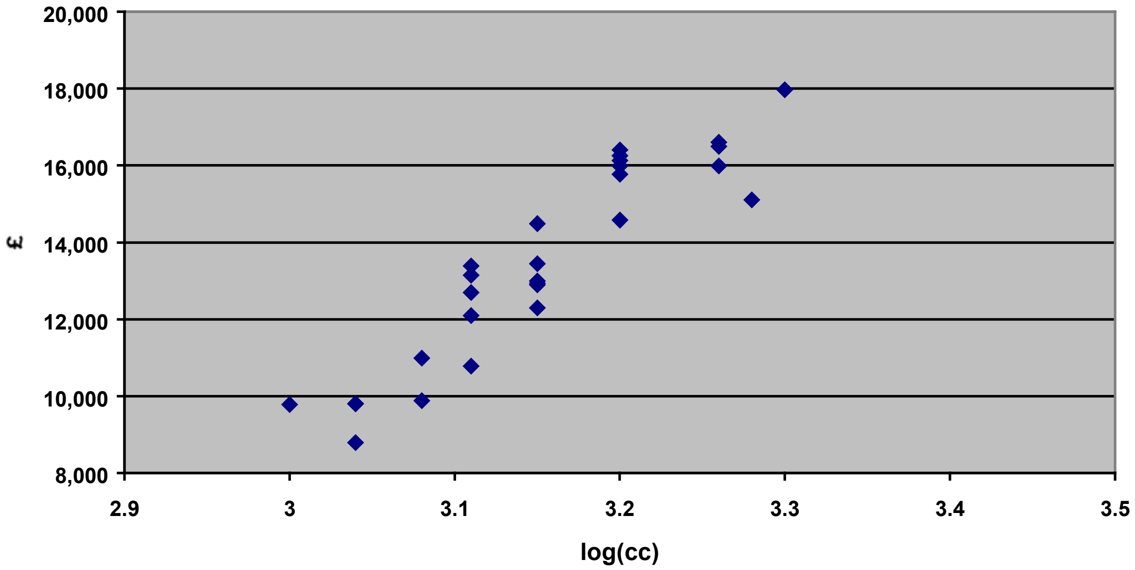
Figure 1 below shows a scatter plot of price (£) versus cylinder capacity (cc) for the data in *Table 1* above.

Figure 1: Price vs Cylinder capicity



The above graph shows a relatively a linear relationship between the two metric variables (price and cc). However to investigate further the relationship between these two variables we can apply the universal method of logarithmic transformation to the *cc variable.* This transformation discounts larger values of *cc* and leaves smaller and intermediate ones intact and has the effect of increasing the linearity of relationship (see graph below).

Figure 2: Price vs log(cc)



1.3 Correlation coefficient

The descriptive statistic most widely used to summarise a relationship between metric variables is a measure of the *degree of linearity* in the relationship. It is called *product-moment correlation coefficient* denoted by the symbol ***r*** and it is defined by

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left[\left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right) \right]$$

where $\overline{x}$ and s_x are the mean and standard deviation of the *x* variable and $\overline{y}$ and s_y are the mean and standard deviation of the *y* variable.

The product moment correlation coefficient has many properties, the most important of which are

1. Its numerical value lies between – *1* and + *1*, inclusive

1. If *r = 1*, then the scatterplot shows that the data lie exactly on a straight line with a positive slope; if *r = -1*, then the scatterplot shows that the data lie on a straight line with a negative slope.

1. An *r = 0* indicates that there is no linear component in the relationship between the two variables.

These properties emphasise the role of *r* as a measure of linearity. Essentially, the more the scatterplot looks like a positively sloping straight line, the closer *r* is to *+1*, and the more the scatterplot looks like a negatively sloping straight line, the closer *r* is to *–1*.

Using the equation above, ***r*** is estimated for the relationship shown in *Fig. 1* to be 0.92 indicating a strong linear relationship between the price of a new car and cylinder capacity. For the relationship shown in *Fig. 2,* ***r*** is estimated to be 0.93, indicating that using the logarithmic transformation does indeed increase the linearity of relationship between the two metric variables. The **LINEST** function in EXCEL was used to estimate *r* in both of the above relationships.

2. Fitting Curves- Regression analysis

In the sections above we showed how to summarise the relationships between metric variables using correlations. Although correlations are valuable tools, they are not powerful enough to handle many complex problems in practice. Correlations have two major limitations:

- They summarise only linearity in relationships.

- They do not yield models for how one variable influences another.

The tool of regression analysis overcomes these limitations by using mathematical curves to summarise relationships among several variables. A ***regression model*** consists of the mathematical curve summarising the relationship together with measures of variation from that curve. Because any type of curve can be used, relationships can be non-linear.

Regression analysis also easily accommodates transformations of variables and categorical variables, and it provides a host of diagnostic statistics that help assess the utility of variables and transformations and the impact of such features as outliers and missing data.

2.1 Models

A model describes *how* a process works. For scientific purposes, the most useful models are statements of the form " *if* certain conditions apply, then certain consequences follow". The simplest of such statements assert that the list of conditions result in a single consequence without fail. For example, we learn in physics that if an object falls toward earth, then it accelerates at about 981 centimetres per second.

A less simple statement is one that assesses a tendency: "Loss in competition tends to arouse anger." While admitting the existence of exceptions, this statement is intended to be universal, that is anger is the expected to loss in competition.

To be useful in documenting the behaviour of processes, models must allow for a range of *consequences* or outcomes. They must also be able to describe a range of *conditions*, fixed levels of predictor variables (x), for it is impossible to hold conditions constant in practice. When a model describes the range of consequences corresponding to a fixed set of conditions, it describes *local* behaviour. A summary of the local behaviours for a range of conditions is called *global* behaviour. Models are most useful if they describe global behaviour over a range of conditions encountered in practice. When they do, they allow us to make predications about the consequences corresponding to conditions that have not actually been observed. In such cases, the models help us reason about processes despite being unable to observe them in complete detail.

2.2 Linear Regression

To illustrate this topic refer back to the sample of cars in *Table 1* and *Figure 1* (the outcome of the scatterplot of *price vs. cc*) above. Our eyes detect a marked linear rend in the plot. Before reading further, use a straight-edge to draw a line through the points that appear to you to be the best description of the trend. Roughly estimate the co-ordinates of two points (not necessarily points corresponding to data points) that lie on the line. From these two estimated points, estimate the slope and y intercept of the line as follows:

Let (x_1,y_1) and (x_2,y_2) denote two points, with $(x_1 \neq y_1)$ on a line whose equation is $y = a + bx$.

Then the slope of the line is

$$b = \frac{\text{Difference in } y \text{ coordinates}}{\text{Difference in } x \text{ coordinates}} = \frac{y_1 - y_2}{x_1 - x_2}$$

and the y intercept is

$$a = \frac{x_1 y_2 - x_2 y_1}{x_1 - x_2}$$

Next, describe the manner in which the data points deviate from your estimated line. Finally, suppose you are told that car has cylinder capacity of 1600cc, and you are asked to use your model to predict the price of the car. Give your best guess at the range of plausible market values. If you do all these things, you will have performed the essential operations of a linear regression analysis of the data.

If you followed the suggestions in the pervious paragraph, you were probably pleased to find that regression analysis is really quite simple. On the other hand, you may not be pleased with the prospect of analysing many large data sets "by eye" or trying to determine a complex model that relates price cylinder capacity, horse power, weight and length simultaneously. To do any but the most rudimentary, the help of a computer is needed.
Statistical software does regression calculations quickly, reliably, and efficiently. In practice one never has to do more than enter data, manipulate data, issue command that ask for calculations and graphs, and interpret output. Consequently, computational formulas are not presented here.

The most widely available routines for regression computations use *least squares* methods. In this section the ideals behind *ordinary least squares* (OLS) are explained. Ordinary least squares fits a curve to data pairs $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ by minimising the sum of the squared vertical distances between the y values and the curve. Ordinary least squares is a fundamental building block of most other fitting methods.

2.3 Fitting a line by ordinary least squares

When a computer program (in this case the LINEST function in EXCEL) is asked to fit a straight-line model to the data in *Figure 1* using the method of ordinary least squares, the following equation is obtained

$$\hat{y} = 307 + 9.11x$$

The symbol y stands for a value of Price (response variable), and the symbol ^ over the y indicates that the model gives only an estimated value. The symbol x (predictor variable) stands for a value of cylinder capacity.

This result can be put into the representation

Observation = Fit and Residual

where y is the observation, $\hat{y}$ is the fit(ted) value and $y - \hat{y}$ is the residual.

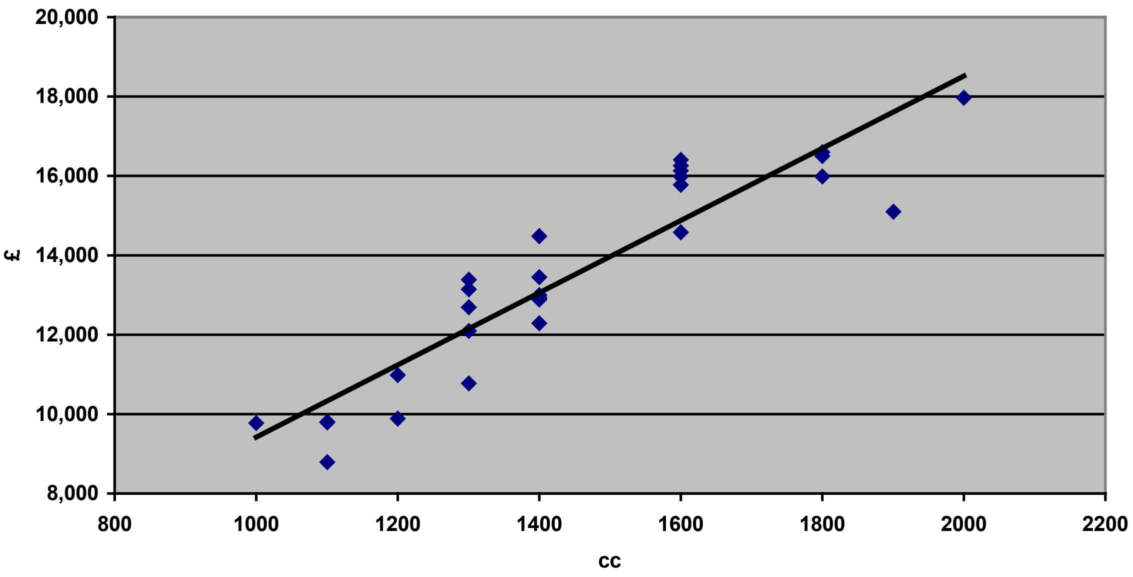
Consider car *No. 1* in *Table 1* that has a cylinder capacity *x = 1300*. The corresponding observed price is y = 13,390. The fitted value given by the ordinary least square line is

$$\begin{aligned}\hat{y} &= 307 + 9.11(1300)\\ &= 307 + 11,843\\ &= 12,150\end{aligned}$$

The vertical distance between the actual price and the fitted price is $y - \hat{y} = 13,390 - 12,150 = 1,240$, which is the residual. The positive sign indicated the actual price is *above* the fitted line. If the sign was negative it means the actual price is *below* the fitted line.

Figure 3 below shows a scatterplot of the data with ordinary least squares line fitted through the points. This plot confirms that the computer can be trained to do the job of fitting a line. *The OLS line was fitted using the linear trend line option in WORD for a chart.*

Figure 3: Price vs Cylinder capicity and OLS line



Another output from the statistical software is a measure of variation: ***s*** = 1028. This measure of variation is the standard deviation of the vertical differences between the data points and the fitted line, that is, the *standard deviation of the residuals*.

An interesting characteristic of the method of least squares is: for any data set, the residuals from fitting a straight line by the method of OLS sum to zero (assuming the model includes a *y* – intercept term). Also because the mean of the OLS residuals is zero, their standard deviation is the square root of the sum of their squares divided by the degrees of freedom. When fitting a straight line by OLS, *the number of degrees of freedom is two less than the number of cases, denoted by n-2* because 1) the residuals sum to zero and 2) the sum of the products of the fitted values and residuals, case by case is zero.

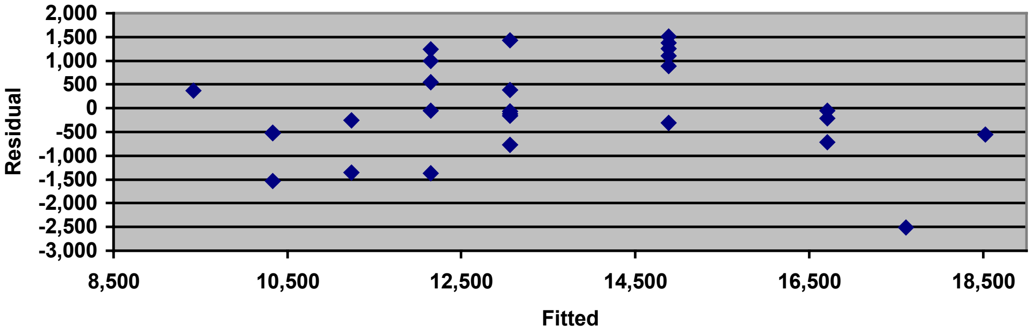
2.4 Analysis of Residuals

Two fundamental tests are applied to residuals from a regression analysis: a test for normality and a scatterplot of residuals versus fitted values. The first test can be performed by checking the percentage of residuals within in one, two and three standard deviations of their mean, which is zero. The second test gives visual cues of model inadequacy.

Table 2. Fitted values and Residuals for new car data

	CC	Price	Fitted	Residual
			Price	
		£		
	(cc)	(p)	(fp)	(res=p-fp)
1	1300	13,390	12,150	1,240
2	1600	15,990	14,883	1,107
3	1300	10,780	12,150	- 1,370
4	1100	9,810	10,328	- 518
5	1600	15,770	14,883	887
6	1300	12,095	12,150	-55
7	1600	16,255	14,883	1,372
8	1400	12,935	13,061	- 126
9	1200	9,885	11,239	- 1,354
10	1600	16,130	14,883	1,247
11	1000	9,780	9,417	363
12	1400	13,445	13,061	384
13	1600	16,400	14,883	1,517
14	1100	8,790	10,328	- 1,538
15	1400	12,995	13,061	- 66
16	1900	15,100	17,616	- 2,516
17	1300	12,700	12,150	550
18	2000	17,970	18,527	- 557
19	1300	13,150	12,150	1,000
20	1800	16,600	16,705	- 45
21	1100	9,795	10,328	- 533
22	1400	12,295	13,061	- 766
23	1800	16,495	16,705	- 210
24	1400	12,895	13,061	- 166
25	1200	10,990	11,239	- 249
26	1800	15,990	16,705	- 715
27	1600	14,575	14,883	- 308
28	1400	14,485	13,061	1,424

Figure 4: Scatterplot of Residual vs. Fitted from new car data



What do we look for in a plot of residuals versus fitted values? We look for a plot that suggests *random scatter*. As we noted above, the residuals satisfy the constraint

$$\sum\left(y-\hat{y}\right)\hat{y} \quad = \quad 0$$

Where the summation is done over all cases in the data set. The constraint, in turn, implies that the product-moment correlation coefficient between the residuals and the fitted values is zero. If the scatterplot is somehow not consistent with this fact because it exhibits a trend or other peculiar behaviour, then we have evidence that the model has not adequately captured the relationship between *x* and *y*. This is the primary purpose of residual analysis: to seek evidence of inadequacy.

The scatterplot in *Figure 3* suggests random scatter and the regression equation above is, therefore, consistent with the above constraint.

3. Hedonic Regression model

3.1 The inclusion of additional metric variables

So far the variable used to account for the variation in the price of a new car is a measure of a physical characteristic which is more or less permanent, though cylinder capacity can change with improvements or deteriorations. This variable does not link up directly with economic factors in the market place, however. Regardless of the cylinder capacity of the car the price of a new is also related to the horsepower, weight, length and number of doors.

Defining the characteristics of a new car as follows

x_1 = cylinder capacity (cc)
x_2 = number of doors (d)
x_3 = horse power (ps)

we propose to fit a model of the form

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \qquad (3.1)$$

When applying regression models (Hedonic regression) to a car index it is usual to fit a semilogarithmic form as it has been proven to fit the data best. That is

$$\log_e \hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \qquad (3.2)$$

This model relates the logarithm of the price of a new car to absolute values of the characteristics. Natural logarithms are used , because in such a model a *b* coefficient, if multiplied by a hundred measures, will provide an estimate of the percentage increase in price due to a one unit change in the particular characteristic or "*quality*" , holding the level of the other characteristics constant.

Using the LINEST function in EXCEL (or PROC REG in SAS) the following estimates for the *b* coefficients are obtained when the above model is applied to the data in *Table 1*.

$$\log_e \hat{y} = 8.43 + 0.000436x_1 + 0.033094x_2 + 0.003605x_3 \qquad (3.3)$$

The interpretation of the above equation is as follows. Keeping the level of other characteristics constant

- A one unit change in cylinder capacity gives a 0.0436% increase in price
- A one unit change in the number of doors gives a 3.3094% increase in price
- A one unit change in brake horsepower gives a 0.3605% in crease in price.

3.2 The inclusion of categorical variables

The next step is to incorporate power steering, ABS system and air bags into the model. The variables are categorical variables: their numeric values *1*, and *0* stand for the inclusion or exclusion of these features in a car.

The semilogarithmic form of the model is now.

$$\log_e \hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 \qquad \textit{(3.4)}$$

where

x_4 = weight $\quad x_7$ = ABS system (abs)
x_5 = length $\quad x_8$ = air bags (ab)
x_6 = power steering (pst)

Using the LINEST function in EXCEL (or PROC REG in SAS) the following estimates for the *b* coefficients are obtained when the above model is applied to the relevant data in *Table 1*.

Equation (3.5)

$$\log_e \hat{y} = 9.37 + 0.000089x_1 + 0.0197x_2 + 0.0023x_3 + 0.0015x_4 - 0.0054x_5 + 0.0649x_6 + 0.113x_7 - 0.0075x_8$$

.

The regression coefficients obtained from *Equation (3.5)* are interpreted as follows. Keeping the level of other characteristics constant

- A one unit change in cylinder capacity gives a 0.0089% increase in price.
- A one unit change in the number of doors gives a 1.97% increase in price.
- A one unit change in brake horse power gives a 0.23.% increase in price.
- A one unit change in weight (kg) gives a 0.15% increase in price.
- A one unit change in the length (cm) gives a 0.54% decrease in price
- The inclusion of power steering gives a 6.49% increase in price.
- The inclusion of an ABS system gives an 11.26% increase in price.
- The inclusion of air bags gives a 0.75% decrease in price

In *Section 4* below it is shown that there is strong collinearity between *weight* and *length* in the above regression model and, therefore, *length* will be omitted from the model.

The regression model now becomes

$$\log_e \hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 \qquad (3.6)$$

where

x_4 = weight
x_5 = power steering (pst)
x_6 = ABS system (abs)
x_7 = air bags (ab)

Using the LINEST function in EXCEL (or PROC REG in SAS) the following estimates for the *b* coefficients are obtained when the above model is applied to the relevant data in *Table 1*.

$$\log_e \hat{y} = 8.43 + 0.00008x_1 + 0.015x_2 + 0.002x_3 + 0.0005x_4 + 0.107x_5 + 0.079x_6 - 0.034x_7 \qquad (3.7)$$

The interpretation of the above equation is as follows. Keeping the level of other characteristics constant

- A one unit change in cylinder capacity gives a 0.008% increase in price.
- A one unit change in the number of doors gives a 1.5% increase in price.
- A one unit change in brake horse power gives a 0.2% increase in price.
- A one unit change in weight (kg) gives a 0.05% increase in price.
- The inclusion of power steering gives a 10.7% increase in price.
- The inclusion of an ABS system gives a 7.9% increase in price.
- The inclusion of an airbag gives a 3.4.% decrease in price

Section 4 below shows that collinearity in not an issue in the regression model described in *Equation (3.7)*.

The output of the regression results for *Equation (3.7)* is displayed below. All the regression coefficients are significantly different from zero with *t* statistics (*t* ratios) greater than *0.8*. An R-square of 96% indicates that almost all of the variation in the price of new cars is explained by the selected predictors.

```
Model: MODEL1
Dependent Variable: P


                                   Analysis of Variance

        Source             DF          Sum of            Mean          F Value          Prob>F
                                      Squares          Square

        Model               7         1.04297         0.14900           71.462          0.0001
        Error              20         0.04170         0.00208
        C Total            27         1.08467
              Root MSE            0.04566              R-square                 0.9616
              Dep Mean            9.49107              Adj R-sq                 0.9481
              C.V.                0.48110


                                 Parameter Estimates

                                  Parameter     Standard Error          T for H0:
Variable             DF           Estimate                           Parameter=0      Prob > |T|

INTERCEP              1           8.425031         0.14860470           56.694          0.0001
CC                    1        0.000077042         0.00009113            0.845          0.4079
D                     1           0.015308         0.01319688            1.160          0.2597
PS                    1           0.002427         0.00073364            3.309          0.0035
W                     1           0.000487         0.00017666            2.758          0.0121
PST                   1           0.106869         0.03691626            2.895          0.0090
ABS                   1           0.079148         0.04351762            1.819          0.0840
AB                    1          -0.033942        -0.02419823            1.403          0.1761


Variable             DF           Variance
                                 Inflation

INTERCEP              1         0.00000000
CC                    1         7.28722659
D                     1         1.45581431
PS                    1         3.01093302
W                     1        10.43520369
PST                   1         2.68458544
ABS                   1         5.83911068
AB                    1         1.92580528
```

As part of any statistical analysis is to stand back and criticise the regression model and its assumptions. This phase is called ***model diagnosis***. If under close scrutiny the assumptions seem to be approximately satisfied and the model can be used to predict and understand the relationship between response and the response and the predictors.

In *Section 5* below the regression model, as described in *Equation (3.7)*, is proven to be adequate for predicting and to understanding the relationship between response and predictors for the new car data described in *Table 1*.

3.3 Classic Definition of Hedonic Regression

As we can see above, the hedonic hypothesis assumes that a commodity (e.g. a new car) can be viewed as a bundle of characteristics or attributes (e.g. cc, horse power, weight, etc.) for which implicit prices can be derived from prices of different versions of the same commodity containing different levels of specific characteristics.

The ability to desegregate a commodity and price its components facilities the construction of price indices and the measurement of price change across versions of the same commodity. A number of issues arise when trying to accomplish this.

1. What are the relevant characteristics of a commodity bundle?
1. How are the implicit (implied) prices to be estimated from the available data?
1. How are the resulting estimates to be used to construct price or quality indices for a particular commodity?
1. What meaning, if any, is to be given to the resulting constructs?
1. What do such indices measure?
1. Under what conditions do they measure it unambiguously?

Much criticism of the hedonic approach has focused on the last two questions, pointing out the restrictive nature of the assumptions required to establish the “existence” and meaning of such indices. However, what the hedonic approach attempts to do is provide a tool for estimating “missing” prices, prices of particular bundles not observed in the base or later periods. It does not pretend to dispose of the question of whether various observed differences are demand or supply determined, how the observed variety of model in the market is generated, and whether the resulting indices have an unambiguous interpretation of their purpose.

4 Collinearity

Suppose that in the car data (see *Table 1*) the car weight in pounds in addition to the car weight in kilograms is used as a predictor variable. Let x_1 denote the weight in kilograms and let x_2 denote the weight in pounds. Now since one kilogram is the same as *2.2046* pounds,

$$\beta_1 \mathrm{x}_1 + \beta_2 x_2 = \beta_1 x_1 + \beta_2 \left(2.2046 x_1\right) = \left(\beta_1 + 2.2046\beta_2\right) x_1 = \gamma x_1$$

with $\gamma = \beta_1 + 2.2046\beta_2$. Here γ represents the "true" regression coefficient associated with the predictor weigh when measured in pounds. Regardless of the value of γ, there are infinitely many different values for β_1 and β_2 that produce the same value for γ. If both x_1 and x_2 are included in the model, then β_1 and β_2 cannot be uniquely defined and cannot be estimated from the data.

The same difficulty occurs if there is a linear relationship among any of the predictor variables. If some set of predictor variables $x_1, x_2 \ldots, x_m$ and some set of constants $c_1, c_2 \ldots, c_{m+1}$ not all zero

$$c_1 x_1 + c_2 x_2 + \ldots + c_m x_m = c_{m+1} \qquad (4.1)$$

for all values of $x_1, x_2 \ldots, x_m$ in the data set, then the predictors $x_1, x_2 \ldots, x_m$ are said to be ***collinear***. Exact collinearity rarely occurs with actual data, but approximate collinearity occurs when predictors are nearly linearly related. As discussed later, approximate collinearity also causes substantial difficulties in regression analysis. Variables are said to be collinear even if Equation *(4.1)* holds only approximately. Setting aside for the moment the assessment of the effects of collinearity, how is it detected?

The search for collinearity between predictor variables is assessed by calculating the correlation coefficients between all pairs of predictor variables and displaying them in a table.

Table 4: Correlation table for predictor variables

	cc	d	ps	w	l	pst	abs
d	0.42754						
ps	0.75827	0.23098					
w	0.90567	0.42623	0.78991				
l	0.91509	0.40526	0.78197	0.98931			
pst	0.59539	0.43901	0.48691	0.67540	0.60361		
abs	0.82684	0.2129	0.63492	0.80978	0.86680	0.34752	
ab	0.55255	0.7412	0.46523	0.52271	0.58908	0.34995	0.64500

The above table of correlations are only between pairs of predictors and cannot assess more complicated (near) linear relationships among several predictors and expressed in Equation *(4.1)*. To do so the *multiple coefficient of determination*, R_j^2, obtained from regressing the *j*th predictor variable on all the other predictor variables is calculated. That is, x_j is temporarily treated as the response in this regression. The closer this R_j^2 is to *1* (or *100%*), the more serious the collinearity problem is with respect to the *j*th predictor.

4.1 Effects on parameter estimates.

The effect of collinearity on the estimates of regression coefficients may be best seen from the expression giving the standard errors of those coefficients. Standard errors give a measure of expected variability for coefficients – the smaller the standard error the better the coefficient tends to be estimated. It may be shown that the standard error of the *j*th coefficient, b_j, is given by

$$se\left(b_j\right) = s\sqrt{\frac{1}{1-R_j^2}\cdot\frac{1}{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2}} \tag{4.2}$$

where, as before, R_j^2 is the R^2 value obtained from regressing the *j*th predictor variable on all other predictors. Equation *(4.2)* shows that, with respect to collinearity, the standard error will be smallest when R_j^2 is zero, that is, the *j*th predictor is not linearly related to the other predictors. Conversely, if R_j^2 is near *1*, then the standard error of b_j is large and the estimate is much more likely to be far from the true value of β_j.

The quantity

$$VIF_j = \frac{1}{1-R_j^2} \tag{4.3}$$

is called the ***variance inflation factor*** (***VIF***). The large the value of *VIF* for a predictor x_j, the more severe the collinearity problem. As a guideline, many authors recommend that a *VIF* greater than *10* suggests a collinearity difficulty worthy of further study. This is equivalent to flagging predictors with R_j^2 grater than 90%.

Table 5 below presents the results of the collinearity diagnostics for the regression model outlined in *Equation 3.5* (using PROC REG in SAS).

Table 5 Variance Inflation Factors (VIP)

cc	7.36028546
d	1.54999525
ps	3.07094954
w	221.98482270
l	246.25396926
pst	4.73427617
abs	7.87431815
ab	3.28577252

From *Tables 4 and 5* above it is obvious that there is a strong liner relationship between the predictor variables *w* and *l* in the regression model in *Equation* (*3.5)* and they are ***collinear***. To over come this collinearity problem the predictor variable *l* (length) will be omitted from the regression model.

4.2 Effects on inference

If collinearity affects parameter estimates and their standard errors then it follows that *t- ratios* will also be affected.

4.3 Effects on prediction

The effect of collinearity on prediction depends on the particular values specified for the predictors. If the relationship among the predictors used in fitting the model are preserved in the predictor values used for prediction, then the predictions will be little affected by collinearity. ON the other hand, if the specified predictor values are contrary to the observed relationships among the predictors in the model, then the predictions will be poor.

4.4 What to do about collinearity

The best defence against the problems associated with collinear predictors is to keep the models as simple as possible. Variables that add little to the usefulness of a regression model should be deleted from the model. When collinearity is detected among variables, none of which can reasonably be deleted from a regression model, avoid extrapolation and beware of inference on individual regression coefficients.

Table 6 below presents the results of the collinearity diagnostics for the regression model outlined in *Equation 3.7* (using PROC REG in SAS).

Table 6 Variance Inflation Factors (VIP)

cc	7.28722659
d	1.45581431
ps	3.01093302
w	10.43520369
pst	2.68458544
abs	5.83911068
ab	1.92580528

Note that the predictor variable *w* (weight) does not have a *VIP* value sufficiently greater than *10* to warrant exclusion from the model. *Tables 4 and 6* above indicate that the regression model as described in *Equation (3.7)* does not have a problem with collinearity among the variables.

5 Model diagnostics

All the regression theory and methods presented above rely to a certain extent on the standard regression assumptions. In particular it was assumed that the data were generated by a process that could be modelled according to

$$y_i \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_{ik} + e_i \qquad \text{for } i = 1,2, \ldots.,n \qquad (5.1)$$

where the error terms $e_1, e_2, \ldots, e_n$ are independent of one another and are each normally distributed with mean *0* and common standard deviation σ . But in any practical situation , assumptions are always in doubt and can only hold approximately at best. The second part of any statistical analysis is to stand back and criticise the model and its assumptions. This phase is frequently called ***model diagnosis***. If under close scrutiny the assumptions seem to be approximately satisfied, then the model can be used to predict and to understand the relationship between response and predictors. Otherwise, ways to improve the model are sought, once more checking the assumptions of the new model. This process is continued until either a satisfactory model is found or it is determined that none of the models are completely satisfactory. Ideally, the adequacy of the model is assessed by checking it with a new set of data. However, that is a rare luxury; most often diagnostics based on the original set must suffice.

The study of diagnostics begins with the important topic of residuals.

5.1 Residuals – standardised residuals

Most of the regression assumptions apply to the error terms $e_1, e_2, \ldots, e_n$. However the error terms cannot be obtained, and the assessment of the errors must be based on the *residuals* obtained as the actual value minus the fitted value that the model predicts with all unknown parameters estimated for the data. Recall that in symbols the *i*th residual is

$$\overset{\wedge}{e_i} \quad = \quad y_i - b_0 - b_1 x_1 - b_2 x_2 - \cdots - b_k x_{ik} \qquad \text{for } i = 1,2, \ldots.,n \qquad (5.2)$$

To analyse residuals (or any other diagnostic statistic), their behaviour when the model assumption *do* hold and, if possible, when at least some of the assumptions *do not* hold must be understood. If the regression assumptions all hold, it my be shown that the residuals have normal distributions with *0* means. It may also be shown that the distribution of the *i*th residual has the standard deviation $\sigma\sqrt{1-h_{ii}}$, where h_{ii} is the ith diagonal element of the "hat matrix" determined by the values of the set of predictor variables. (See Appendix I), but the particular formula given there is not needed here. In the simple case of a single predictor model it may be shown that

$$h_{ii} = \frac{1}{n} + \frac{\left(x_i - \overline{x}\right)^2}{\sum_{j=1}^{n}\left(x_j - \overline{x}\right)^2} \qquad (5.3)$$

Note in particular that the standard deviation of the distribution of the *i*th residual is not σ , the standard deviation of the distribution of the *i*th error term e_i . It may be shown that , in general,

$$\frac{1}{n} \quad \le \quad h_{ii} \le \quad 1 \qquad (5.4)$$

so that

$$0 \quad \le \quad \sigma\sqrt{1-h_{ii}} \quad \le \quad \sigma\sqrt{1-\frac{1}{n}} \quad \le \quad \sigma \qquad (5.5)$$

It may be seen from *Equation (5.3)* and also argued in the general case that h_{ii} is at its minimum value, *1/n*, when the predictors are all equal to their mean values. On the other hand, h_{ii} approaches its maximum value, *1*, when the predictors are very far from their mean values. Thus residuals obtained from data points that are far from the centre of the data set will tend to be smaller than the corresponding error terms. Curves fit by least squares will usually fit better at extreme values for the predicators than in the central part of the data.

Table 3 below, displays the h_{ii} values (along with many other diagnostic statistics that will be discussed) for the regression of log (price) on the seven predicators described above (cc, d, ps, w, l, pst, abs and ab).

Table 3: Diagnostic Statistics for Regression Model

	Dep Var	**Predict**		**Standard**	**t-**	**Hat Diag**	**Cook's**
Obs	**P**	**Value**	**Residual**	**Residual**	**Residual**	h_{ii}	**D**
1	9.5000	9.4673	0.03270	0.772	0.7644	0.1408	0.012
2	9.6800	9.6865	-0.00647	-0.156	-0.1518	0.1704	0.001
3	9.2900	9.2477	0.04230	1.207	1.2213	0.4099	0.126
4	9.1900	9.1546	0.03540	0.981	0.9799	0.3767	0.073
5	9.6700	9.6622	0.00780	0.188	0.1833	0.1739	0.001
6	9.4000	9.4753	-0.07530	-1.814	-1.9346	0.1730	0.086
7	9.7000	9.6775	0.02250	0.554	0.5441	0.2094	0.010
8	9.4700	9.4466	0.02340	0.571	0.5615	0.1974	0.010
9	9.2000	9.2328	-0.03280	-1.022	-1.0232	0.5059	0.134
10	9.6900	9.6865	0.00353	0.085	0.0827	0.1704	0.000
11	9.1900	9.1663	0.02370	0.608	0.5978	0.2712	0.017
12	9.5100	9.5122	-0.00216	-0.053	-0.0512	0.1870	0.000
13	9.7100	9.6865	0.02350	0.566	0.5558	0.1704	0.008
14	9.0800	9.1886	-0.10860	-2.706	-3.3131	0.2279	0.270
15	9.4700	9.4466	0.02340	0.571	0.5615	0.1974	0.010
16	9.6200	9.6222	-0.00220	-0.083	-0.0809	0.6615	0.002
17	9.4500	9.4447	0.00528	0.136	0.1321	0.2707	0.001
18	9.8000	9.7690	0.03100	0.877	0.8714	0.4005	0.064
19	9.4800	9.4237	0.05630	1.389	1.4241	0.2106	0.064
20	9.7200	9.7536	-0.03360	-0.830	-0.8228	0.2132	0.023
21	9.1900	9.1828	0.00721	0.207	0.2021	0.4189	0.004
22	9.4200	9.4677	-0.04770	-1.125	-1.1329	0.1367	0.025
23	9.7100	9.7310	-0.02100	-0.503	-0.4938	0.1646	0.006
24	9.4600	9.4864	-0.02640	-0.725	-0.7160	0.3618	0.037
25	9.3000	9.2998	0.000171	0.006	0.0055	0.5679	0.000
26	9.6800	9.7051	-0.02510	-0.592	-0.5821	0.1409	0.007
27	9.5900	9.6226	-0.03260	-1.302	-1.3267	0.6988	0.492
28	9.5800	9.5041	0.07590	1.826	1.9500	0.1723	0.087

To compensate for the differences in dispersion among the distributions of the different residuals, it is usually better to consider the ***standardised residuals*** defined by

$$i\text{th standardised residual} = \frac{\hat{e}}{s\sqrt{1-h_{ii}}} \qquad \text{for } i = 1, 2, \ldots, n \qquad (5.6)$$

Notice that the unknown σ has been estimated by s. If n is large and if the regression assumptions are all approximately satisfied, then the standardised residuals should behave about like standard normal variables. *Table 3* also lists the residuals and standardised residuals for all 28 observations.

Even if all the regression assumptions are met, the residuals (and the standardised residuals) are not independent. For example, the residuals for a model that includes an intercept term always add to zero. This alone implies they are negatively correlated. It may be shown that, in fact, the theoretical correlation coefficient between the ith and jth residuals (or standard residuals) is

$$\frac{-h_{ij}}{\sqrt{(1-h_{ii})(i-h_{jj})}} \qquad (5.7)$$

where h_{ij} is the ijth element of the hat matrix. Again, the general formula for these elements is not needed here. For the simple single-predictor case it may be shown that

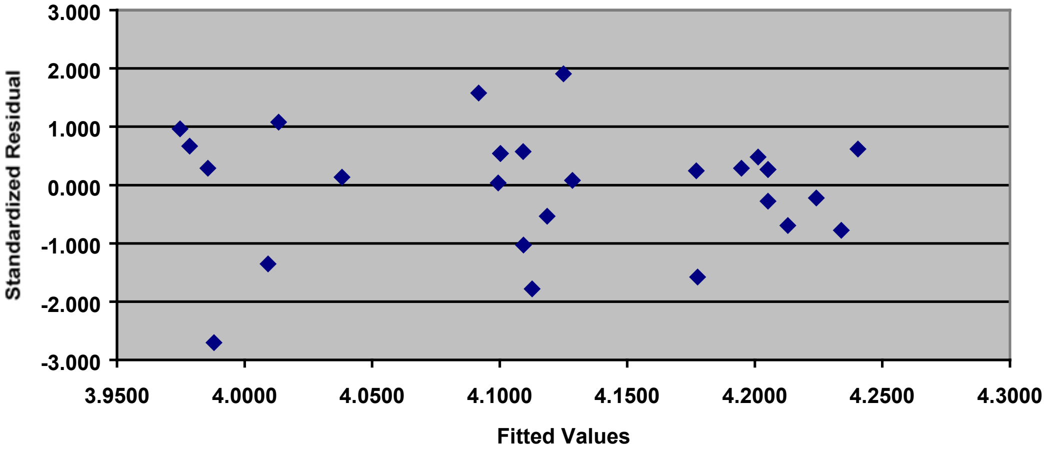
$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (5.8)$$

From Equations (5.3), (5.7) and (5.8) (and in general) we see that the correlations will be small except for small data sets and/or residuals associated with data points very far from the central part of the predictor values. From a practical point of view this small correlation can usually be ignored, and the assumptions on the error terms can be assessed by comparing the properties of the standardised residuals to those of independent, standard normal variables.

5.2 Residual Plots

Plots of the standardised residuals against other variables are very useful in detecting departures from the standard regression assumptions. Many of the most common problems may be seen by plotting (standardised) residuals against the corresponding fitted values. In this plot, residuals associated with approximately equal-sized fitted values are visually grouped together. In this way it is relatively easy to see if mostly negative (or mostly positive) residuals are associated with the largest and smallest fitted values. Such a plot would indicate curvature that the chosen regression curve did not capture. *Figure 5* displays the plot of Standardised Residuals versus fitted values for the new car data.

Figure 5: Scatterplot of Standardized Residual vs. Fitted from new car data



The residual plot in *Figure 5* above displays a mixture of positives and negative and thus shows no general inadequacies.

Another important use for the plot of residuals versus fitted values is to detect lack of common standard deviation among different error terms. Contrary to the assumption of common standard deviation, it is not uncommon for variability to increase as the values for response variables increase. This situation does not occur for the data contained in *Figure 5*.

5.3 Outliers

In regression analysis the model is assumed to be appropriate for all the observations. However, it is not unusual for one or two cases to be inconsistent with the general pattern of the data in one way or another. When a single predictor is used such cases may be easily spotted in the scatterplot data. When several predictors are employed such cases will be much more difficult to detect. The non conforming data points are usually called ***outliers***. Sometimes it is possible to retrace the steps leading to the suspect data point and isolate the reason for the outlier. For example, it could be the result of a recording error, If this is the case the data can be corrected. At other times the outlier may be due to a response obtained when variables not measured were quite different than when the rest of the data were obtained. Regardless of the reason for the outlier, its effect on regression analysis can be substantial.

Ouliers that have unusual response values are the focus here. Unusual responses should be detectable by looking for unusual residuals preferably by checking for unusually large standardised residuals. If the normality of the error terms is not in question, then a standard residual larger than *3* in magnitude certainly is unusual and the corresponding case should be investigated for a special cause for this value.

5.3.1 Studentized Residuals (*t* – residuals)

A difficulty with looking at standardised residuals is that an outlier, if present, will also affect the estimate of σ that enters into the denominator of the standardised residual. Typically, an outlier will inflate *s* and thus deflate the standardised residual and mask the outlier. One way to circumvent this problem is to estimate the value of σ use in calculating the *i*th standard residual using all the data except the *i*th case. Let $s_{(i)}$ denote such an estimate where the subscript *(i)* indicates that the *i*th case has been deleted. This leads to the ***studentized residual*** defined by

$$i\text{th studentized residual} = \frac{\hat{e_i}}{s_{(i)}\sqrt{1-h_{ii}}} \qquad \text{for } i = 1, 2, \ldots, n \qquad (5.6)$$

The next question to be asked is "how do these diagnostic methods work the new car data?". *Table 3* lists diagnostic statistics for the regression model as applied to the new car data. Notice that there is only case (*observation No. 14*) where the standardised and studentized residuals have a significant difference and where the standard or studentized residuals are above *3* in magnitude. These results indicate that, in general, there are no outlier problems associated with the regression model described in *(3.5)* above.

5.4 Influential Observations

The principle of ordinary least squares gives equal weight to each case. On the other hand, each case does not have the same effect on the fitted regression curve. For example, observations with extreme predictor values can have substantial influence on the regression analysis. A number of diagnostic statistics have been invented to quantify the amount of influence (or at least potential influence) that individual cases have in a regression analysis. The first measure of influence is provided by the diagonal elements of the *hat matrix*.

5.4.1 Leverage

When considering the influence of individual cases on regression analysis, the *i*th diagonal element of the hat matrix h_{ii} is often called the ***leverage*** for the *i*th case, which means a measure of the *i*th data point's influence in a regression with respect to the predicator variables. In what sense does h_{ii} measure influence? It may be shown that $\hat{y}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij} y_j$ so that $\delta \hat{y}_i / \delta y_i = h_{ii}$, that is h_{ii} is the rate of change of the *i*th fitted value with respect to the *i*th response value. If h_{ii} is small, then a small change in the *i*th response results in a small change in the corresponding fitted value. However, if h_{ii} is large, then a small change in the *i*th response produces a large change in the corresponding $\hat{y}_i$.

Further interpretation of h_{ii} as leverage is based on the discussion in *Section 5.1*. There is was shown that the standard deviation of the sampling distribution of the *i*th residual is not σ but $\sigma\sqrt{1-h_{ii}}$. Furthermore, h_{ii} is equal to its smallest value *1/n*, when all the predicators are equal to their mean values. These are the values for the predictors that have the least influence on the regression curve and imply, in general, the largest residuals. On the other hand, if the predictors are far from their means, then h_{ii} approaches its largest value of *1* and the standard deviation of such residuals is quite small. In turn this implies a tendency for small residuals, and the regression curve is pulled toward these influential observations.

How large might a leverage value be before a case is considered to have large influence? It may be shown algebraically that the average leverage over all cases is *(k+1)/n*, that is,

$$\frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{k+1}{n} \qquad (5.10)$$

where *k* is the number of predictors in the model. On the basis of this result, many authors suggest making cases as influential if their leverage exceeds two or three time *(k+1)/n*.

For the new car data displayed in *Table 3* and using the regression model as described in *Equation (3.7)* we estimate;

1. *k = 7*
2. *(k+1)/n* = 8/28 = 0.2857
3. 2 *x (k+1)/n* = 0.5714
4. 3 *x (k+1)/n* = 0.8571

In *Table 3* only two observations (No.'s 16 and 27) are above the 2 *x (k+1)/n* threshold and none of the observations are above the 3 *x (k+1)/n* threshold. This result indicates that there are no observations with extreme predictors value impacting on the slope of the fitted values and hence none of the observation have undue influence on the regression results.

5.5 Cook's Distance

As good as large leverage values are in detecting cases influential on the regression analysis, this criterion is not without faults. Leverage values are completely determined by the values of the predictor variables and do not involve the response values at all. A data point that possesses large leverage but also lies close to the trend of the other data will not have undue influence on the regression results.

Several statistics have been proposed to better measure the influence of individual cases. One of the most popular is called ***Cook's Distance***, which is a measure of a data point's influence on regression results that considers both the predictor variables and the response variables. The basic idea is to compare the predictions of the model when the *i*th case is and is not included in the calculations.

In particular, Cook's Distance, D_i, for the *i*th case is defined to be

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{(k+1)s^2} \qquad (5.11)$$

where $\hat{y}_{j(i)}$ is the predicted or fitted value for case *j* using the regression curve obtained when *i* is omitted. Large values of D_i indicate that case i has large influence on the regression results, as then $\hat{y}_j$ and $\hat{y}_{j(i)}$ differ substantially for many cases. The deletion of a case with a large value of D_i will alter conclusions substantially. If D_i is not large, regression results will not change dramatically even if the leverage for the *i*th case is large. In general, if the largest value of D_i is substantially less than *1*, then no cases are especially influential. On the other hand, cases with D_i greater than *1* should certainly be investigated further to more carefully assess their influence on the regression analysis results.

In *Table 3* only one observation *No. 27* has the largest value of Cook's Distance, 0.722 and leverage value, 0.6988. However, neither value is high enough to influence the regression analysis results.

What is next once influential observations have been detected? If the influential observation is due to incorrect recording of the data point, an attempt to correct that observation should be made and the regression analysis rerun. If the data point is known to be faulty but cannot be corrected, then that observation should be excluded for the data set. If it is determined that the influential data point is indeed accurate, it is likely that the proposed regression model is not appropriate for the problem at hand. Perhaps an important predictor variable has been neglected or the form of the regression curve is not adequate.

5.6 Transformations

So far a variety of methods for detecting the failure of some of the underlying assumptions of regression analysis have bee discussed. Transformations of the data, either of the response and/or the predictor variables, provide a powerful method for turning marginally useful regression models into quire valuable models in which the assumptions are much more credible and hence the predictions much more reliable. Some of the most common and most useful transformations include logarithms, square roots, and reciprocals. Careful consideration of various transformations for data can clarify and simplify the structure of relationships among variables.

Sometimes transformations occur "naturally" in the ordinary reporting of data. As an example, consider a bicycle computer that displays , among other things, the current speed of the bicycle in miles per hour. What is really measured is the time it takes for each revolution of the wheel. Since the exact circumference of the tire is stored in the computer, the reported speed is calculated as a constant divided by the measured time per revolution of the wheel. The speed reported is basically a reciprocal transformation of the measured variable.

As a second example, consider petrol consumption in a car. Usually these values are reported in miles per gallon. However, they are obtained by measuring the fuel consumption on a test drive of fixed distance. Miles per gallon are then calculated by computing the reciprocal of the gallons per mile figure.

A very common transformation is the ***logarithm transformation***. It may be shown that a logarithm transformation will tend to correct the problem of non constant standard deviation in case the standard deviation of e_i is proportional to the mean of y_i. If the mean of y doubles, then so does the standard deviation of e and so forth.

6. Sampling Distributions and Significance testing

6.1 Introduction

This section discusses the standardisation of the sample mean. Introducing notation, let $y_1, y_2 \cdots, y_n$ denote a sample of *n* numbers, let $\overline{y}$ denote the mean of the sample, and let *s* denote the sample standard deviation. Assume the sample is from a process or from a very large population so that in either case the question of a finite population correction can be ignored. The long-run process or population mean is denoted by μ, which is the theoretical mean. Then the sample standard error of the sample mean is

$$\frac{s}{\sqrt{n}}$$

and the standardised mean is

$$t = \frac{\overline{y} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\sqrt{n}\left(\overline{y} - \mu\right)}{s} \qquad (6.1)$$

This is the difference between the sample and theoretical means divided by the sample standard error. Usual statistical notation for this quantity is the letter *t*, and the quantity is known as the *t* statistic. Many statistics are referred to as *t* statistics because the idea of dividing the difference between a sample and theoretical quantity by a sample standard error is pervasive in statistical applications.

6.2 t - distribution

If we let $y_1, y_2 \cdots, y_n$ denote a sample of size n drawn randomly from a normal distribution with mean μ and standard deviation σ. Let $\overline{y}$ and s denote the sample mean and standard deviation. Then the sampling distribution of the *t* statistics defined in the equation above is the *t* distribution with *n-1* degrees of freedom (denoted by *v*).

A well know theorem indicates that

1. A *t* distribution with *n-1* degrees is symmetric and mound shaped.
2. Provided *n-1* ≥ 3, the standard deviation of the *t* distribution is $\sqrt{(n-1)/(n-3)}$.
3. When *n-1=1*, the mean and standard deviation of the *t* distribution does not exist.
4. When *n-1=2*, the mean exists but the standard deviation does not.
5. As *n* grows large without bound (i.e. *n > 30*), the *t* distribution converges to the standard normal distribution.

6.3 Application to significance testing

Significance testing is a process of probabilistic inference that uses sampling distributions to compare behaviour in data with theories about the process that generated the data. Consider a situation in which data from a process in statistical control (i.e. the various dimension of a process are within in acceptable limits) whose cross-sectional distribution is normal are drawn, the value of the long-run process mean, denoted by μ, is in doubt, and the value of the long-run process standard deviation σ is not known. One way to approach inference about μ is to venture a guess, called a theory or hypothesis, about the value of μ. After the data are collected, the value of the guess is compared with the value of the sample mean.

Because sample means vary from sample to sample a criterion for determining whether a specific sample mean deviates from the guess by more than an amount that can be attributed to natural sampling variation is needed. The *t* statistic in equation *(6.1)* above and its associated *t* distribution with *n-1* degrees of freedom provide such criterion.

The numerical value of the guess is called the ***null hypothesis*** and is denoted by H_0. If μ_0 denotes the numerical guess at μ, then $H_0 : \mu = \mu_0$ defines the null hypothesis. Once the null hypothesis is defined, the notation H_0 is used to refer to it.

To conduct a test of significance, a ***test statistic*** that forms a comparative link between some function of the data and the long-run process mean μ is defined. The analyst must be able to state the sampling distribution of the test statistic when the null hypothesis is assumed to be true. Saying that the null hypothesis is true means that a "good guess" has been made, that is, that μ_0 and the actual long-run value of μ coincide.

The test statistic must be constructed so that if a good guess is not made, the statistic sends an appropriate signal, which is exactly what the *t* statistic does. The logic of this is as follows. Because the sampling distribution of the *t* statistic is known when H_0 is true, an interval of values expected to be observed, called the interval of plausible values, can be constructed. Now if after the data are collected and the value of the *t* statistic

$$t = \frac{\bar{y}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{\sqrt{n}\left(\bar{y}-\mu_0\right)}{s} \qquad (6.2)$$

is computed, and the *t* statistic falls outside the interval of plausible values, then there is a reason to suspect that the hypothesis was not really a good guess. Notice that the value of *t* is obtained by dividing the difference between the sample mean obtained from the data and the null hypothesis value μ_0 by the sample standard error of the mean.

When the value of the *t* statistic computed for actual data falls outside the interval of plausible values, it has fallen into the ***critical region*** of the test and the value is statistically significant. If the analyst believes the signal the test gives and concludes that the hypothesis is not a good guess, then the ***null hypothesis is rejected***. The implication of this language is that if the actual value of the *t* statistic falls in the interval of plausible values, then the null hypothesis is not rejected.

The interval of plausible values plays a fundamental role. Values of the *t* statistic not in this interval are deemed to be "critical" and to signal rejection of the null hypothesis. The interval of plausible values is chosen to make it unlikely that the test statistic rejects the null hypothesis when it is true. More formally, the interval of plausible values is chosen so that when the null hypothesis is true, an acceptably small proportion of the possible *t* statistics falls outside the interval, according to the sampling *t* distribution with *n-1* degrees of freedom.

Critical regions are estimated using statistical tables know as ***t – tables***. Using these tables it is possible to estimate that when the null hypothesis was assumed to be true, 99% of the possible *t* statistic values were between – 4.604 and 4.604, and the interval between these values is used as the interval of plausible values. The probability of rejecting a true null hypothesis was therefore only 0.01 (or 1%). Since only one out of a hundred possible *t* statistic values would lead to rejecting a true hull hypothesis, the analyst would feel confident that the test of significance is not misleading. Put another way, if a *t* statistic value outside the interval of plausible values (and therefore inside the critical region) is observed, this justifies doubts about the truth of the null hypothesis. The 1% probability that the *t* statistic falls in the critical region is called the ***significance level*** of the test. It is a measure of the risk of incorrectly concluding that the null hypothesis is false.

The risk of rejecting a true null hypothesis is only one kind of risk. Another is the risk of not rejecting a false null hypothesis. It is a trade-off between these two risks that forces analysts to use nonzero significance levels. A test that never rejects a null hypothesis cannot signal that the guess at μ is no good. Analysts run some risk of rejecting a true null hypothesis to discover a false one. Such is the trade-off inherent n trying to discover new truth from imperfect or incomplete data.

There is no completely objective method for choosing the significance level of a significance test. The prevailing practice is to choose significance levels more or less by convention, the most common choices being 10%, 5% and 1%. The smaller the significance level, the larger the interval of plausible values, and the larger the *t* statistic has to be, in absolute value, to fall in the critical region and signal rejection of the null hypothesis.

Three types of errors are possible in significance testing

1. ***Type I***, this is the error of rejecting a true null hypothesis. This means that the significance level is the probability of committing a Type I error.
2. ***Type II***, this is the error in not rejecting a false null hypothesis.
3. ***Type III***, this refers to answering an irrelevant question. In formulating problems analysts usually try to define the problem in terms that make it easy to solve. Ding this creates the risk of "defining away" the real problem, that is, setting up a problem that can be solved but whose salient features do not match the real problem.

6.4 CHI-SQUARE statistics

Pearsons X^2 statistic is a measure of association (summarising relationships between categorical variables) for multi-way tables. Pearson's statistic is often referred to as a *chi-squared statistic*. *Chi-square* is a transliteration of the mathematical symbol χ^2, which is the Greek letter *chi* to the second power. This notation is used to stand for the family of mathematical curves that describe the sampling distribution of Pearson's statistic under certain conditions. In particular the **chi-squared distribution** is the approximate sampling distribution of Pearsons X^2 statistic when the null hypothesis of no association is true. Below we discuss the connection between χ^2 distributions and sample variances of samples from the normal population.

For a two way table (i.e. a 2 x 2) the following notation is used

		W		
		W_1	W_2	*Total*
	V_1	a	c	$a + c = n_{1.}$
V	V_2	b	d	$b + d = n_{2.}$
	Total	$a + b = n_{.1}$	$c + d = n_{.2}$	$a + b + c + d = n$

The two categorical variables are *V* and *W*, the counts in each cell of the cross classification are denoted by *a, b, c* and *d*, and the row, column, and grand totals are denoted by the *n*'s with appropriate subscripts. The value of Pearson's chi-squared statistic, denoted by X^2 , is given by the formula

$$\mathrm{X}^2 = \frac{n(ad-bc)^2}{(n_{.1})(n_{.2})(n_{1.})(n_{2.})} \qquad (6.2)$$

It is to be noted that when there is association among the categorical variables the values of the X^2 tend to be larger, on the whole, than when there is not association among the variables. This principle is the basis for a significance test in which the null hypothesis is that the categorical do not interact in the universe. When the null hypothesis is true and the table is *2 x 2*, then the sampling distribution of X^2 is approximately χ^2 with *1* degree of freedom. In general an *r x c* table has degrees of freedom *v=(r-1)(c-1).* We know that if the variables do interact in the universe the X^2 values will be large and so sufficiently large values of X^2 should be taken as evidence against the null hypothesis.

It can be shown theoretically that the 95th percentile of the χ^2 distribution with 1 degree of freedom is 3.841. If a 5% significance level is desired, the null hypothesis is rejected whenever the value of X^2 is grater that 3.841. If this rule is followed, there is a 5% risk of declaring a true null hypothesis false. It can also be shown that the 99th percentile of the χ^2 distribution with 1 degree of freedom is 6.635, so if a 1% significance level is desired, the null hypothesis is rejected whenever the value of X^2 is greater than 6.635. This information is read from a chi-squared distribution tables.

The mathematical theory for Pearson's chi-squared statistic says that as the sample size , *n*, gets larger, the sampling distribution of X^2 becomes more nearly like a χ^2 distribution with 1 degree of freedom, provided the null hypothesis of no association is true. This type of statement also occurs in the central limit effect, which guarantees approximate normality of totals and means, provided the sample size is large enough. A rough guideline is "do not rely on the χ^2 approximation unless the sample size is at least 50 and all the cell frequencies are at least 5.

Another thorny question in applications of Pearson's chi-squared statistic is that of sampling design. The theory discussed above assumes simple random sampling with replacement, but in practice this design is rare. Research on complex sampling designs show clearly that Pearson's X^2 statistic has different sampling distributions when different sampling designs are used, and the differences seriously affect the significance levels of significance tests. Because of this, X^2 should be used with cautiously when making probabilistic inferences.

6.5 F statistics

F statistics are used when probabilistic inferences about sources of variation are made. These inferences are called analysis of variance. An analysis of variance is the output from a regression command.

In mathematical theory of statistics a random quantity has an ***F distribution*** with v_1 and v_2 degrees of freedom if it is a ratio of two independent chi-squared random quantities divided by their degrees of freedom. In symbols

$$F = \frac{U_1/v_1}{U_2/v_2}$$

where U_1 and U_2 are independent, v_1 has a chi-squared distribution with v_1 degrees of freedom and U_2 has a chi-squared distribution with v_2 degrees of freedom. The parameter v_1 is called the numerator degrees of freedom; v_2 the denominator degrees of freedom.

The mean and standard deviation of the *F* distribution with v_1 and v_2 degrees of freedom are

$$\mu = \frac{v_2}{v_2-2} \text{ and } \sigma = \sqrt{\frac{2v_2^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}}$$

Note that the mean does not exist if v_2 is less than or equal to 2, and the standard deviation does not exist if v_2 is less than or equal to 4.

Tables of *F* distributions are complicated because they must display distributions for each possible combination of the numerator and denominator degrees of freedom. Access to computer software is essential for practical uses of the *F* distribution. There is no use limiting yourself to the percentiles shown in the typical tables.

Appendix I - The Matrix Approach to Multiple Regression

The multiple regression model that has been written as the *n* equations

$$y_i \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_{ik} + e_i \qquad \text{for } i = 1,2, \ldots,n \qquad (A1)$$

may also be expressed very economically in vector-matrix notation. First define the column vectors

$$\underset{(n\times 1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \underset{[(k+1)\times 1]}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \underset{(n\times 1)}{\mathbf{e}} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

and the matrix

$$\underset{(n\times(k+1))}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

Then recalling the definition of matrix multiplication, *Equations (A1)* may be written compactly as

$$\mathbf{y} \quad = \quad \mathbf{X}\beta + \mathbf{e} \qquad (A2)$$

The principle of ordinary least squares says to estimate the components of β by minimising the quantity

$$\mathbf{S}(\beta) = \sum_{i=1}^{n} \left[y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik} \right]^2 = (\mathbf{y} - \mathbf{X}\beta)^{'} (\mathbf{y} - \mathbf{X}\beta) \qquad (A3)$$

This may be accomplished by solving the system of *k+1* linear equations obtained from computing the partial derivatives and setting

$$\frac{\delta}{\delta\beta}\mathbf{S}(\beta) \quad = \quad 0$$

This in turn yields the so-called *normal equation*

$$\mathbf{X}^{'}\mathbf{X}\beta \quad = \quad \mathbf{X}^{'}\mathbf{y} \qquad (A4)$$

Here $X^{'}$ denotes the transpose of matrix *X*.

A proof using algebra (but not calculus) that a solution of the normal equations provides the least squares estimates may be obtained as follows: Let *b* be any solution of the *normal equations (A4*). Then

$$\begin{aligned}\mathbf{S}(\beta) &= (\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)\\ &= \left[(\mathbf{y}-\mathbf{Xb})-\mathbf{X}(\beta-\mathbf{b})\right]'\left[(\mathbf{y}-\mathbf{Xb})-\mathbf{X}(\beta-\mathbf{b})\right]\\ &= (\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})+\left[\mathbf{X}(\beta-\mathbf{b})\right]'\left[\mathbf{X}(\beta-\mathbf{b})\right]\\ &\quad +(\mathbf{y}-\mathbf{Xb})'\mathbf{X}(\beta-\mathbf{b})+\left[\mathbf{X}(\beta-\mathbf{b})\right]'(\mathbf{y}-\mathbf{Xb})\end{aligned}$$

But since b satisfies the *normal equations (A4)* it is easy to see that the final two “cross-products” terms are each zero. Thus we have the identity

$$\mathbf{S}(\beta) = (\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})+\left[\mathbf{X}(\beta-\mathbf{b})\right]'\left[\mathbf{X}(\beta-\mathbf{b})\right] \qquad \textit{(A5)}$$

The first term on the right hand side of *Equation (A5)* does not involve β ; the second term is the sum of squares of the elements of the vector $\mathbf{X}(\beta-\mathbf{b})$. This sum of squares can never be negative and is clearly smallest (namely zero) when $\beta=\mathbf{b}$. Thus a solution to the normal equations will provide ordinary least squares estimates of the components of β.

If the *(k+1) x (k+1)* dimensional matrix $\mathbf{X}'\mathbf{X}$ is invertible, then *Equation (A4)* has a unique solution which may be written as

$$\mathbf{b} \;=\; \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \qquad \textit{(A6)}$$

The column vector of *fitted values* is then

$$\overset{\wedge}{\mathbf{y}} \;=\; \mathbf{Xb} \;=\; \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \qquad \textit{(A7)}$$

and the column vector of *residuals* is

$$\overset{\wedge}{\mathbf{e}} \;=\; \mathbf{y}-\overset{\wedge}{\mathbf{y}} \;=\; \left[\mathbf{1}-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{y} \qquad \textit{(A8)}$$

By direct calculation it is easy to see that the matrix $\mathbf{H}=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ has the special property $\mathbf{H}'\mathbf{H}=\mathbf{H}$ so that **H** is an *idempotent matrix*. It may also be argued that **H** is a symmetric matrix so that $\mathbf{H}'=\mathbf{H}$. The matrix **H** is sometimes called the *hat matrix* since the observation vector **y** is pre-multiplied by **H** to produce $\overset{\wedge}{\mathbf{y}}$ (**y** hat). It is easy to show that **1-H** is also symmetric and idempotent.

The estimate of σ is then

$$s \;=\; \sqrt{\frac{\overset{\wedge}{\mathbf{e}}'\overset{\wedge}{\mathbf{e}}}{n-k-1}} \;=\; \sqrt{\frac{\left(\mathbf{y}-\overset{\wedge}{\mathbf{y}}\right)'\left(\mathbf{y}-\overset{\wedge}{\mathbf{y}}\right)}{n-k-1}} \;=\; \sqrt{\frac{\mathbf{y}'(\mathbf{1}-\mathbf{H})\mathbf{y}}{n-k-1}} \qquad \textit{(A9)}$$

with *n-k-1* degrees of freedom.

Under the usual regression assumptions, it may be shown that the individual regression coefficient, b_i , has a normal distribution with mean β_i . The standard deviation of the distribution of b_i is given by σ

times the square root of the *i*th diagonal element of the matrix $\left(X^{'}X\right)^{-1}$. The standard error of b_i is obtained similarly by replacing σ by *s*. That is

$$se\left(b_i\right) \quad = \quad s\sqrt{\left[\left(\mathbf{X}^{'}\mathbf{X}\right)^{-1}\right]_{ii}} \qquad\qquad \textit{(A10)}$$

Let $\mathbf{x}^{*} = \left(1, x_1^{*}, x_2^{*}, \cdots, x_k^{*}\right)$ be a row vector containing specific values for the *k* predictor variables for which we want to predict a future value for the response, y^{*} . The prediction is given by $\mathbf{x}^{*}\mathbf{b}$ and the prediction error is $y^{*} = \mathbf{x}^{*}\mathbf{b}$. The standard deviation of the prediction error can be shown to be

$$\sigma_{y^{*}-\mathbf{x}^{*}\mathbf{b}} \quad = \quad \sigma\sqrt{1+\mathbf{x}^{*}\left(\mathbf{X}^{'}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{x}^{*'}} \qquad\qquad \textit{(A11)}$$

The prediction standard error, denoted *predse*, is obtained by replacing σ by *s* in *Equation (A11).* That is

$$predse \quad = \quad s\sqrt{1+\mathbf{x}^{*}\left(\mathbf{X}^{'}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{x}^{*'}} \qquad\qquad \textit{(A12)}$$

Finally, the breakdown of the total sum of squares may be expressed as

$$\left(\mathbf{y}-\overline{\mathbf{y}}\right)^{'}\left(\mathbf{y}-\overline{\mathbf{y}}\right) = \left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right)^{'}\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right) + \left(\mathbf{y}-\hat{\mathbf{y}}\right)^{'}\left(\mathbf{y}-\hat{\mathbf{y}}\right) \qquad\qquad \textit{(A13)}$$

[Total SS = Regression SS + Residual SS]

with degrees of freedom *n-1, k* and *n-k-1*, respectively. Here $\overline{\mathbf{y}}$ is a column vector with the mean $\overline{y}$ in all positions.

References

Griliches, Z. (1961). "Hedonic price indexes for automobiles: An econometric analysis of quality change." In *The Price Statistics of the Federal Government*, General Series, No. 73, pp 137-196. National Bureau of Economic Research, New York

Griliches, Z (1971) . Hedonic Price Indexes Revisited, p 3-15 in Griliches, Zvi (edt). *Price Indexes and Quality Change"*. Harvard University Press, Cambridge, Massachusetts, 1971.

Lancaster, K (1966). A New Approach to Consumer Theory, *Journal of Political Economy* 74, 132-157.