Chapter 5 : Regression

This chapter follows up on some of the ideas from chapter 4.

Here the data is of the same format, that is pairs (x, y), or triples (x_1, x_2, x_3) or even in larger sets or dimensions.

Now the emphasis is treat or think of the variable x as something that is used to predict the outcome y. What this means in more detail is discussed later, so prediction is not yet precisely defined.

In experiments or studies one observes $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Sometimes the experimenter can specify x, other times they are observed in pairs.

Regression, more specifically linear (sometimes called simple linear) regression looks for a straight line so that we model or predict the outcome y in terms of x by the relation

$$y = a + bx$$

When there is more than a single x to *predict* the outcome y the method is called multiple regression. In this chapter we only consider the case of a single x to *predict* the outcome y.

As an example recall the car mileage versus weight. One of our plots was a scatter plot with a straight line overlaying the plot. The *regression* technique is to use the weight of the vehicle (the x value) to *predict* the gas consumption measured by miles per gallon (the outcome y). This is shown in Figure 1.

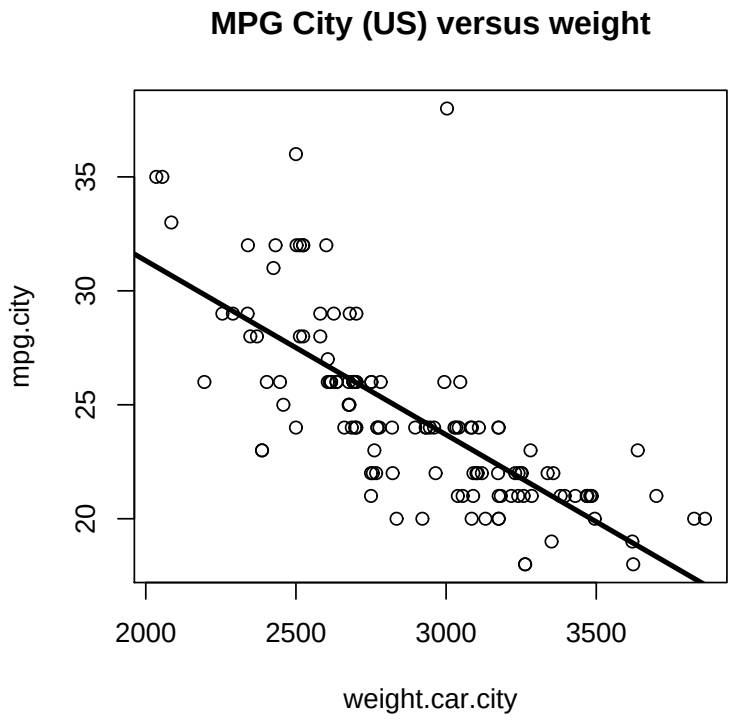

Figure 1: MPG City all 4 Cylinder Cars, Normal Power and Best Line : mpg = 46.6 - 0.0076*weight

First we need to review and consider two things.

What is a straight line?

How is this line found?

A straight line is represented by an equation of the form

$$y = a + bx$$

a is the intercept term, that is the value of y when $x = 0$. The point $(0, a)$ lines on the straight line.

b is the slope. It measures the amount of change in y when x by 1 (unit).

For a given x the value of y is given by $a + bx$

For a given $x + 1$, y takes the value $a + b(x + 1) = a + bx + b$, that is y changes by an amount b.

To see why this is so

$$\begin{aligned} &\{a + b(x + 1)\} - \{a + bx\} \\ &= a + b * x + b - a - b * x \\ &= b \end{aligned}$$

The symbol $*$ here represents multiplication (for ease of reading).

If x by amount dx then y changes by an amount $b * dx$

For a given $x + dx$, y takes the value $a + b(x + dx) = a + b * x + b * dx$, then y changes by an amount

$$\begin{aligned} & \{a + b(x + dx)\} - \{a + bx\} \\ &= a + b * x + b * dx - a - b * x \\ &= b * dx \end{aligned}$$

Any straight line can be determined or described by these two *parameters*, that is by the two values a = intercept and b = slope. This says that these two *parameters* are all the information that is needed to specify and hence draw the straight line.

Regression line

The regression line is chosen to fit as closely as possible to the given data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. This is done by what is known as the *method of least squares*. This method give a recipe for calculating the fitted or estimated intercept and slope. These are given by the formulae

$$\begin{aligned} b &= r\frac{s_y}{s_x} \\ a &= \bar{y} - b\bar{x} \end{aligned}$$

Properties about regression

1. the line $y = a + bx$ is the best fitting line in the sense of *the method of least squares*

   $b = r\frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$

2. the slope b measures the amount y changes on average when x changes by 1 unit

3. if x changes by 1 standard deviation, that is by amount s_x, then y changes by an amount $b * s_x$ or equivalently by an amount r times the standard deviation of y. This last part comes from the property of the regression line slope b. For those interested in the algebra for this we have

   For those who wish to follow the algebra this fact is determined by

$$\begin{aligned} & \{a + b(x + s_x)\} - \{a + bx\} \\ = \;& a + bx + bs_x - a - bx \\ = \;& bs_x \\ = \;& r\frac{s_y}{s_x}s_x \\ = \;& rs_y \end{aligned}$$

   This derivation will not be part of exams.

4. $$r^2 = \frac{\text{variation in } \hat{y} \text{ as } x \text{ pulls it along the line}}{\text{total variation in obxerved values of } y}$$

This says that r^2 measures how much of the variation of the observed y is explained by the regression line in terms of the input or predictor variable x

Revisit mpg versus weight 2004 US cars example. We use only the 4 cylinder non hybrid cars. There are 127 cars or vehicles in the data set, as it contains cars, pickup trucks, SUV, etc. The term car is used generically here to mean motor vehicle.

The first 5 data points are

mpgcity	mpghwy	wt (weight)
28	34	2370
28	34	2348
26	37	2617
26	37	2676
26	37	2617

When we fit the city mpg regressed against weight we use the pairs (wt, mpgcity).

When we fit the highway mpg regressed against weight we use the pairs (wt, mpghwy).

Next we consider the regression of city mpg against weight, that is fit a line

$$\text{mpgcity} = a + b * \text{weight}$$

or in terms of $y = \text{mpgcity}$ and $x = \text{weight}$

$$y = a + b * x$$

Some summary data for the city mpg are

r	-0.744
$\bar{y}$	24.520
$\bar{x}$	2890.748
s_y	16.030
s_x	151937.6

Calculate $b = r * \frac{s_y}{s_x} = -0.744 * \frac{16.030}{151937.6} = -0.00785$ Note this is note quite the slope calculated -0.00763 by the software due to round off error in the numbers recorded in table above.

Using only normal engine 4 cylinder cars, we found in the Chapter 4 examples that $r = -0.74$. Thus $r^2 = .55$, and hence the regression line in terms of weight explains 55% or equivalently a fraction 0.55 of the variation in gas mileage (miles per gallon) of non-hybrid 4 cylinder cars. The weight of the vehicle explains over half of the variability in fuel consumption amongst the 4 cylinder vehicles.

Another very useful term in regression is *residual*

Recall we have n pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Using these we fit the simple linear regression model

$$y = a + bx$$

The fitted slope and intercept are given by the formula

$$\begin{aligned} b &= r\frac{s_y}{s_x} \\ a &= \bar{y} - b\bar{x} \end{aligned}$$

For the given (x_i, y_i) pair the *regression line* predicts that the observed y_i should be

$$\hat{y}_i = a + bx_i$$

The notation $\hat{y}$ (read y hat) is to denote the regression line predicted value of y.

Example :

y = mpgcity	wt (weight)	$\hat{y}$	residual = $y - \hat{y}$
28	2370	28.49712	-0.497 = 28 - 28.497
28	2348	28.665	-.665
26	2617	26.611	-.611
26	2676	26.160	-.160
26	2617	26.611	-.611

Notice that any vehicle with the same weight has the same predicted miles per gallon city fuel consumption, based on this simple linear regression model. Amongst other factors not accounted for in the model are horse power engine output.

Figure 2 shows the first 20 residuals. On this plot we have

- the observed and fitted values (shown as circles) for the first 20 car weights.
- the regression line. Notice that the fitted values of course are on this line.
- the dashed lines show the residuals. Some of these are not appearing entirely on the plot so Figure 3 shows the same plot on a different y axis scale.

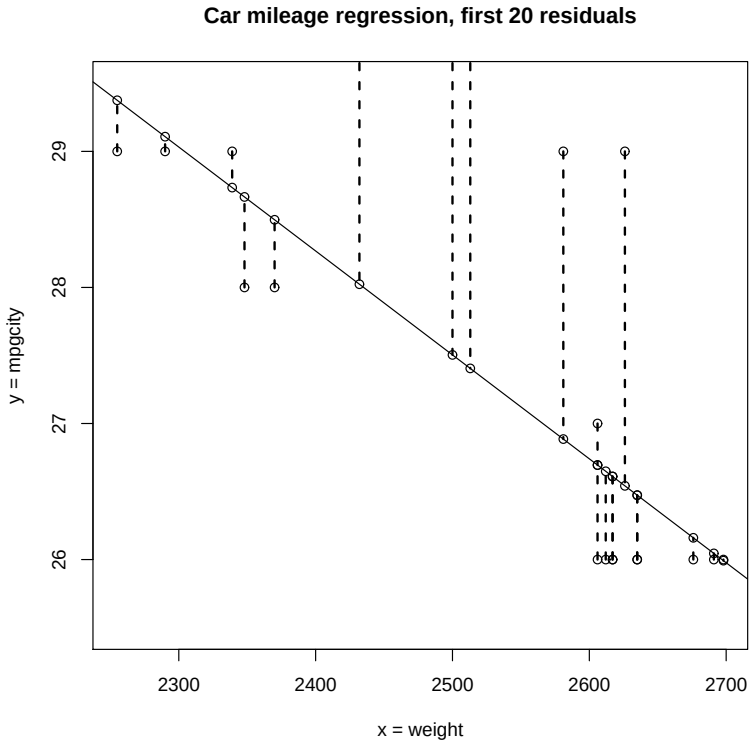

Figure 2: Residuals for first 20 data points

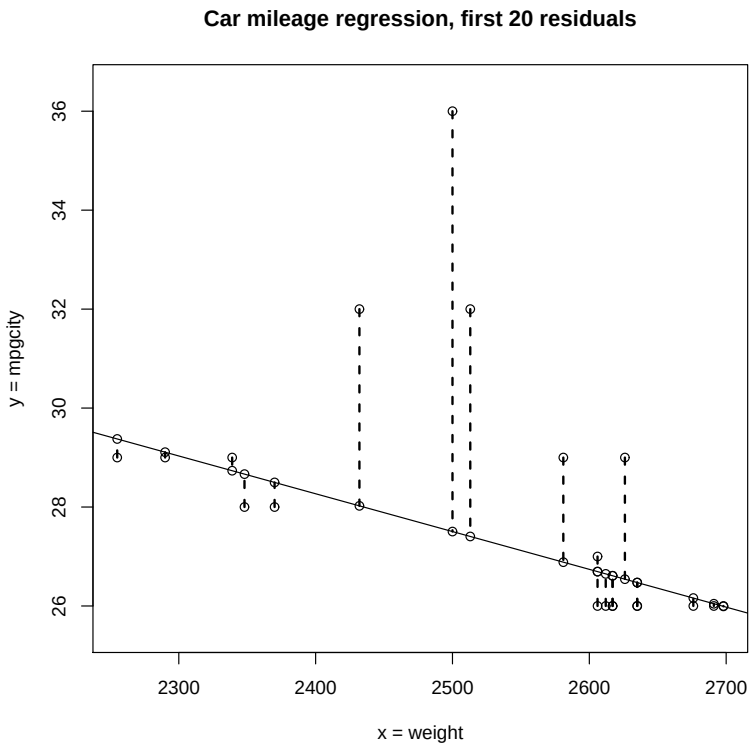

Figure 3: Residuals for first 20 data points (different y-scale

$r^2 = .55$ measures how well the regression line predicts the observed y using the regression line $a + b * x$.

We next examine the residuals a little more. The residuals are centred at 0.

In the next plots the variable resid.car is the car city mpg residuals. This name is chosen to be unique in the software I have used, and has no significance other than a useful name.

Figures 4 and 5 show the scatter plot of residuals against car weights and the histogram of residuals respectively. We notice that a couple of residuals are quite different from the others. They take values around 8 and 15. Except for these 2 residuals, all the others range from around -5 to +5. Are these vehicles different in some quite significant way compared with the others? If so maybe they should be treated as special. It may also be that the data is not recorded correctly.

Going back to the original data it turns out these a special diesel engine, and they are indeed different from the normal gasoline internal combustion engine. In order to get a better predictor for the so called normal engines we remove these two data points and redo the regression analysis.

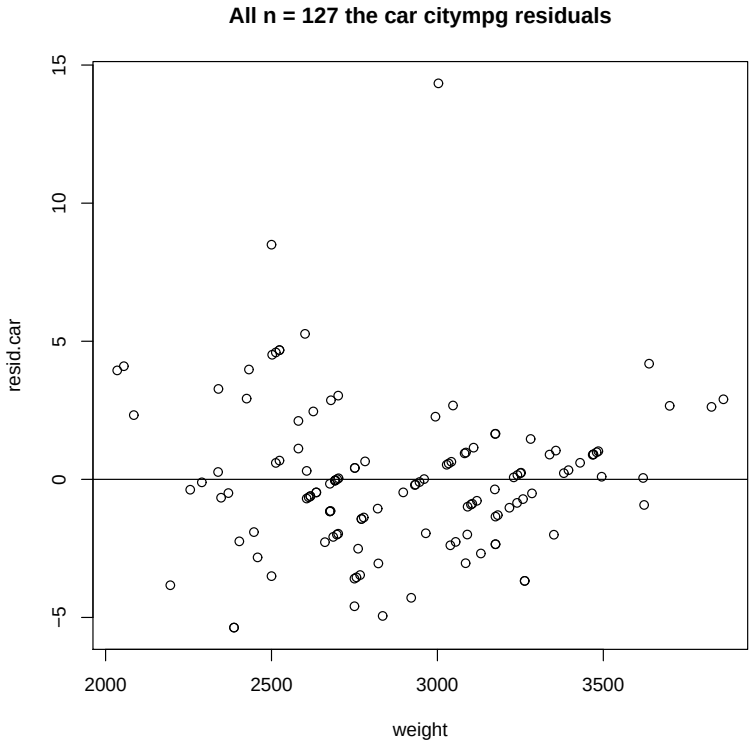

Figure 4: Four Cylinder Car non hybrid Residuals for all data points

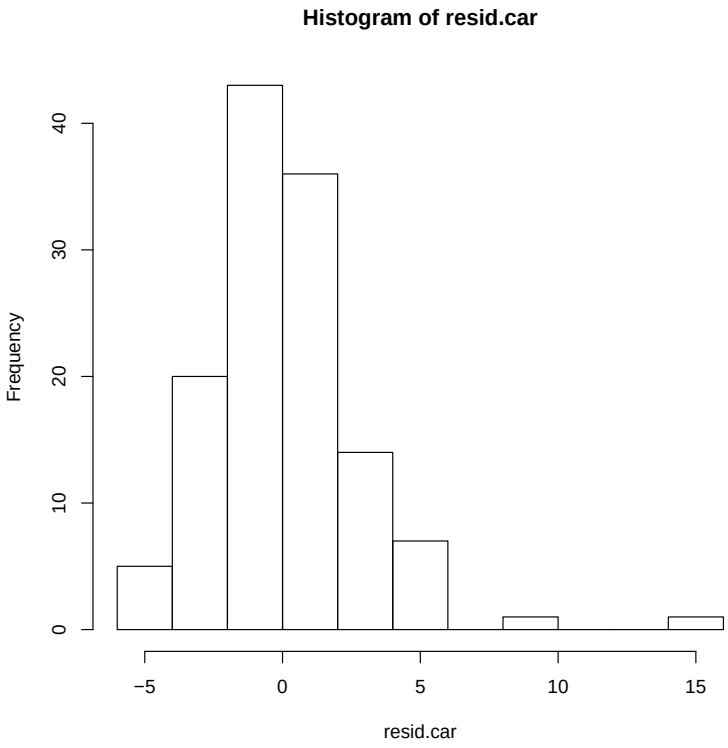

Figure 5: Residuals for all non hybrid 4 cylinder car data points

Aside :

Graphical tools such as scatter plots and histograms help us to identify unusual data points. These are sometimes called *outliers*. Outliers may be due to (i) recording errors, (ii) the data point is correct but quite different in some important feature.

It is reason (ii) that we found in our car mileage and weight data analysis.

The modified data set now contains $n = 125$ data points.

Using these data the fitted regression model is now

$$y = 46.148 - 0.00754 * x$$

The correlation between the *normal engine* cars and weight is

$$r = -0.80$$

The regression model now explains $.64 = (-.8)^2$ proportion (or 64%) of the variation in car city mileage.

Figure 6 shows the relative frequency histogram for the residuals from this regression line (or regression model). Some summary information for the residuals are

mean residual	0
variance residual	5.006
standard deviation residual	2.237

In Figure 6 we also see the residuals for this model are approximately symmetric and bell shaped. On this plot is drawn the normal curve with mean 0 and standard deviation 2.24 (actually 2.237). We see it is quite a good approximation to the residual relative frequency histogram or distribution.

Aside : As discussed in Chapter 2 and 3, it is often the case that the normal distribution is a good description or approximation to the distribution of many types of data, especially from a measurement type of data. Sometimes one needs to use a model to *predict* the mean or average behaviour first. In fact that is what the regression models actually do; they *predict* the mean

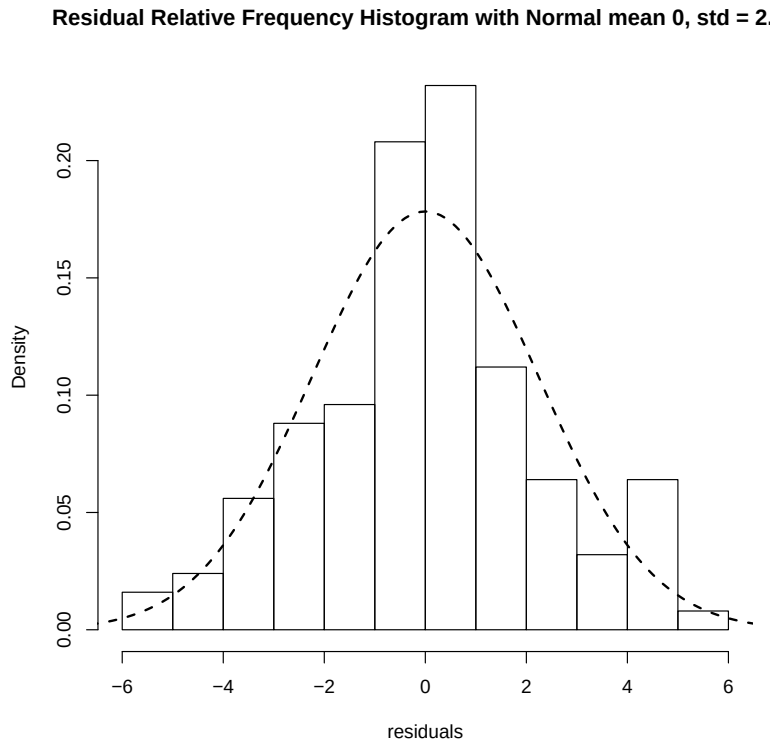

Figure 6: Relative Frequencey Histogram of Residuals for all normal engine 4 cylinder car data points

in this case for a given value of the regressor variable (the car weight in the example above).

As another example consider the data given in Problem 4.42 (p 121, of the Fifth Edition)

It considers measuring the sun's brightness in a given city. This data is immediately after some pollution controls were implemented. It is interesting to learn if these have any measurable effect over the next decade or so. One thought is that as the atmospheric haze decreases the brightness level will correspondingly increase. Thus it becomes interesting to know if the brightness level does increase as year (time) increases. If a linear regression model is reasonable, and the slope (coefficient b) is positive, this would be supporting evidence that the pollution controls are having a measurable effect.

For this data some relevant summary information is given below. The extra labels x and y are to help you remember that you are regressing y against x (Watts regressed against Year). The sample size is $n = 11$.

x	mean Year	1997
	variance Year	11.00
y	mean Watts	249.291
	variance Watts	6.929
	correlation(Year, Watts)	.773

From this information we can calculate the regression line

$$y = -976.141 + 0.614 * x$$

or equivalently written as

$$\text{Watts} = -976.141 + 0.614 * \text{Year}$$

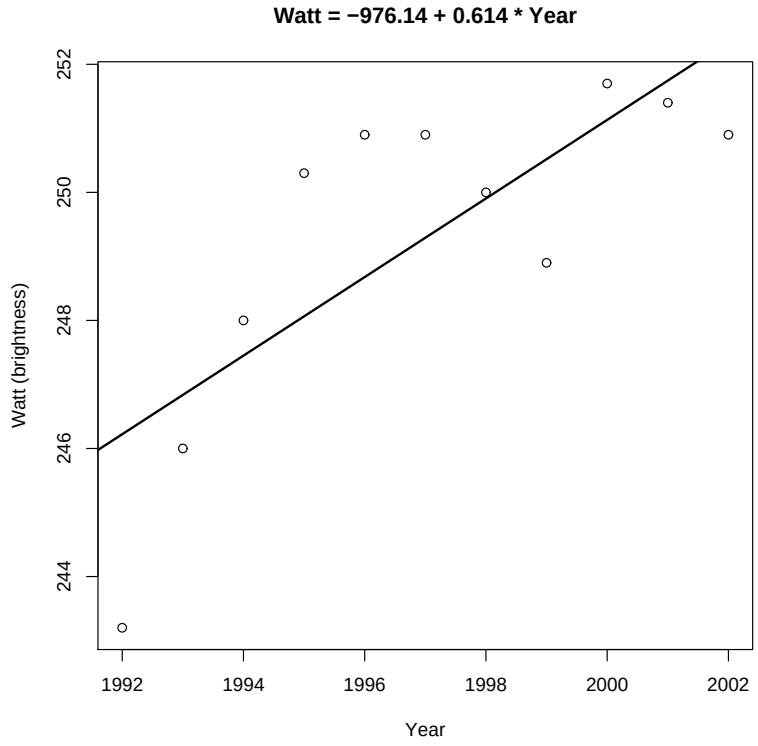

Figure 7: Watts versus Year with Fitted Regression Line

Figure 10 shows the data and regression line. $r^2 = .60$ so this time trend explains 60% of the year to year variation in sun's brightness at this site. This regression line is interesting

Figure **??** shows the residuals. There is nothing to indicate any unusual behaviour in the residuals. Thus the regression model seems reasonable.

Aside There are more formal ways of studying the appropriate fit of models, but these are not considered in this course.

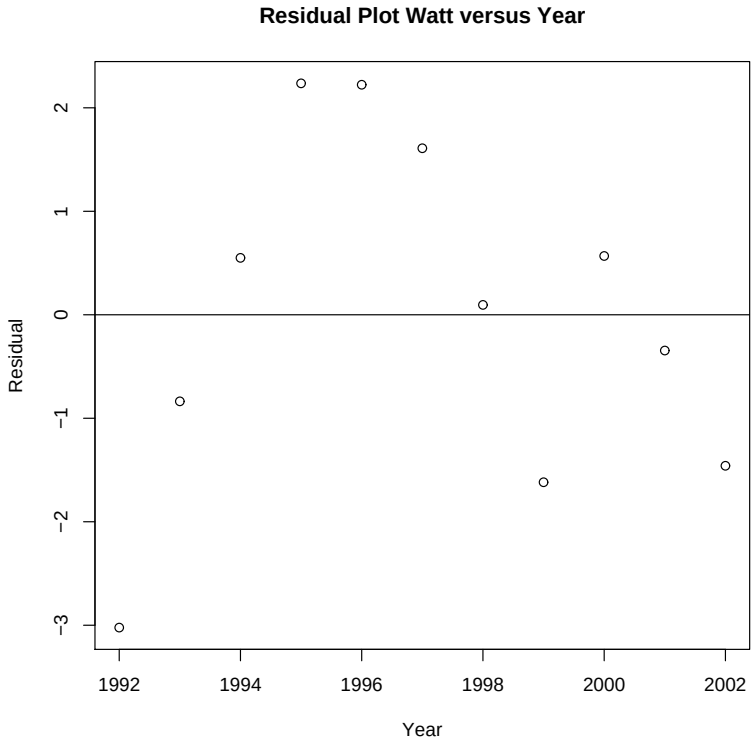

Figure 8: Residuals of Brightness Regression Fit

How can one study whether annual temperature is increasing on this planet? We could make a time series plot of (average annual global) temperature versus year, and also regress temperature against year if the simple linear model is appropriate. In the same idea as above we are interested in knowing if the trend (measured by the slope in the simple linear regression model) is upwards (is the slope positive). There are also many underlying questions of a scientific nature to decide what is causing the increase. The body of scientific evidence is that it is caused by human intervention in terms of atmospheric pollutants.

In Chapter 4 we looked at a diamond price data set. There $r = .989$, which is very close to 1, indicating that the data fall nearly along a straight line. Thus the linear regression model will give a very good prediction of price in terms of carrot size of the diamond. In particular $r^2 = .978$ so that the regression model accounts for 97.8% of the variability in the price. Notice that the variance of the diamond prices is much bigger than the variance of the residuals.

variance diamond price	45643.23
variance residuals	992.25

If you know the carrot size of the diamond you will then be able to guess within about $Singapore 50 of the actual price. The actual prices range from a minimum of $Singapore 223 to a maximum of $Singapore 1086.

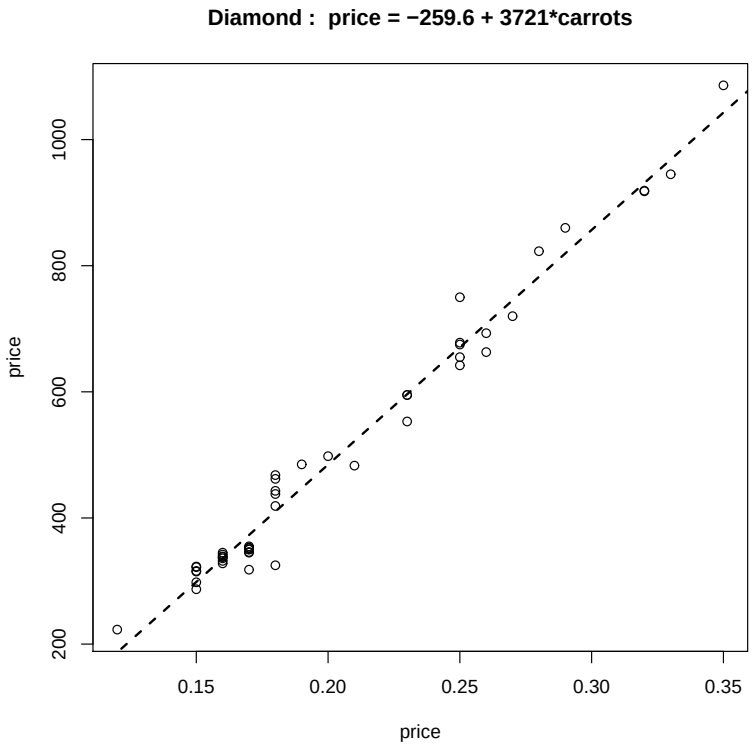

Figure 9: Diamond Price versus Carrots Scatter Plot and Regression

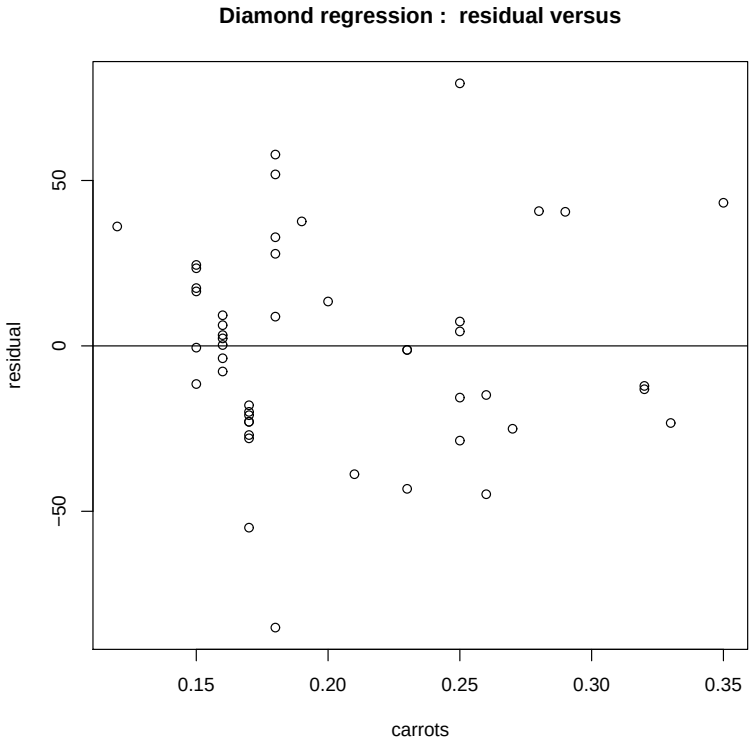

Figure 10: Diamond Regression Residuals versus Carrots