The simple linear regression model has:

$$y_i = b_0 + b_1 x_i + \text{error}_i = \mu_{Y|x} + \text{error}_i.$$

- The term $\mu_{Y|x}$ is there to emphasize that the model says the mean of the y values for a given x is modeled by a line – that depends on x – with slope b_1 and intercept b_0.
- The assumptions about the error terms are that they are a *random sample* from a normal population with a mean of 0 and standard deviation σ.

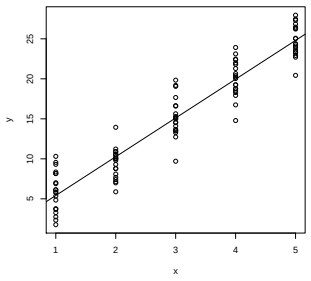

Figure 1: For each x the y values are normally distributed around the line. That is the error terms are normally distributed with mean 0.

In the following we want to investigate three things:

1. How can we test the assumption that the linear model is appropriate?
2. How can we test if the errors are normally distributed and a *random sample* when we don't actually have the errors?
3. How sensitive is the model fitting – finding $\hat{b}_1$ and $\hat{b}_0$ to outliers?

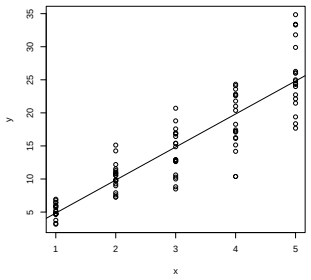

Figure 2: Error terms are normal – but the variance is not constant

Linear model assumption

We will use a data set on `diamonds` from an article at JSE `http://www.amstat.org/publications/jse/v9n2/datasets.chu.html` by Singfat Chu.

```
> f <- "http://www.math.csi.cuny.edu/verzani/classes/MTH214/Computer/diamonds.txt"
> diamonds <- read.table(f)
```

This data set has many variables. The ones we are interested in are for now are `carat`, `price` and `lnprice`.

The price is a worth and carat a measure of size. We should have that larger diamonds are worth more, so these two should be positively correlated.

Question 0.1. Are they? Find the correlation between the two variables. (the function is `cor`.) You can refer to them through `diamonds$carat` and `diamonds$price`.

Question 0.2. Now make a scatterplot of the relationship with `price` as the response variable and `carat` as the predictor. Use a command like:

```
> plot(price ~ carat, data = diamonds)
```

Mentally draw a trend line.

1. Is it increasing?
2. Is it linear?
3. If not, what kind of curve or shape does your trend line have.

Question 0.3. You can add a linear regression fit to your graph with the command

```
> res <- lm(price ~ carat, data = diamonds)
> abline(res)
```

Do so. Does this help you answer your last question better?

To assess a relationship as to whether it is linear from a regression line is pretty easy to do, but to really look to see if the data has a linear trend, versus some other, one should remove the trend and then look. The residual plot will do this. The residual plot looks at the residuals only plotted against the *fitted values* – that is the $\hat{y}_i$ values that fall on the line. To make this plot we have:

```
> plot(fitted(res), resid(res))
```

Or better still, using one of R's diagnostic plots for regression:

```
> plot(res, which = 1)
```

[R has several (6) diagnostic plots to test the assumptions of the model. The value of 1 selects the residuals versus fitted graphic. The value 2 will make a quantile plot of the residuals]

Question 0.4. Make the last of these plots (using `which=1`). The graph should show a dashed line at 0 and a curved line. The dashed line indicates that the fitted values were stripped away so *on average* the residuals are 0. However, the red line tracks any trend left in the residuals.

1. Does the red line show a non-linear trend in the residuals?

Question 0.5. Repeat the above with variable `lnprice` as a response and `carat` as a predictor. The variable `lnprice` is the logarithm of the price. Taking the logarithm is a common practice to turn exponential relationships into linear ones.

1. First make the scatter plot. Does the linear model seem to fit better?
2. Now make a residual plot. Does the linear model seem any better now?

Normality assumption

The assumptions about the model trend are that it is linear. If that is satisfied, one must also check that the assumptions about the error terms are also satisfied. Recall we assume that the error terms are:

1. Normally distributed with mean 0 and standard deviation σ.
2. Are independent

The residuals are our surrogate for the error terms which are unknown as they depend on knowing the true model (b_o and b_1, not $\hat{b}_0$ and $\hat{b}_1$).

To check the the residuals are normally distributed we can make a histogram, quantile-quantile plot or boxplot.

The residuals are given from the model fit by the extractor function `resid`.

```
> x <- resid(res)
> head(x)
```

```
       1        2        3        4        5        6
120.6924 328.6924 328.6924  78.6924 343.7036 257.7036
```

The graphics can then be made with these.

Question 0.6. For the model of `price` by `carat` make the three three plots above for the residuals and discuss. For example, to find the quantile-quantile plot we have:

```
> res <- lm(price ~ carat, data = diamonds)
> qqnorm(resid(res))
```

Do the residuals look normally distributed? Why or why not? Be specific how each graphic shows this.

Question 0.7. Repeat the above for the model of `lnprice` versus `carat`. Do these residuals look normally distributed? Why or why not.

The residuals always have mean 0. (Sample mean). This can be checked:

```
> mean(resid(res))

[1] -5.064491e-15
```

which up to rounding error is 0.

However, the residuals may not indicate that the assumption that each error term has the same standard deviation σ. That is, our model assumes our mean response depends on x, but the variance of the error does not. The term *homoscedasticy* applies when it is the case.

Question 0.8. Make a residual plot of the residuals for the model `price` by `carat`. Does it appear that the assumption on σ is satisfied?

Question 0.9. When the residuals show a violation of the assumption that σ does not depend on x, a transformation is often applied to the response variable. The variable `lnprice` is the result of taking the log-transform. Make the appropriate graph to check if this transformation is consistent with the assumption on a value of σ that does not depend on x.

The issue of independence is also important. Sometimes in the process of data collection the values collected near each other may influence the other values. For example, a really large value may make one extra careful to make sure the next value is observed properly. As such, this next measurement is influenced by the previous which means the results are not independent.

A common graphic to check for this is to plot the residuals against the *lagged* residuals. R makes this easy, provided you use its subscripting to shift the residuals. Here is how.

```
> res <- lm(price ~ carat, data = diamonds)
> x <- resid(res)
> plot(x[-1] ~ x[-length(x)], main = "Plot of previous, versus next")
```

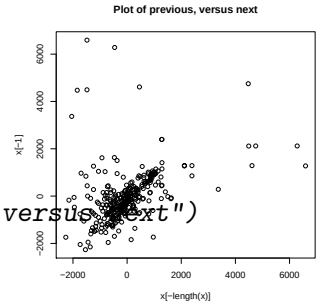

Figure 3: Lag plot (previous versus next) of the residuals for the price by carat size. The lack of a pattern indicates no relationship between subsequent measurements.

This graph for this data is just scattered since the observations are independent – that is the previous value does not influence the next value. It will will sometimes show a pattern if the observations are not independent (it can be non-independent in different ways).

Question 0.10. Make the plot above only for the model of `lnprice` versus `carat`. Does it look like the residuals are independent by this measure?

Question 0.11. Define `x` values as follows – so that the values are related – and then make the plot above. Do you see a pattern?

```
> x <- sin(1:100)
```

Influential observations

The estimates for the regression parameters b_1 and b_0 are sensitive to outliers. When a point causes a big change in the slope of regrssion line, we say it is an influential point. These points are marked by some of the diagnostic plots that R produces. To see, we look at the model restricted to the values where `D==1`.

```
> diam <- subset(diamonds, subset = D == 1)
> res <- lm(price ~ carat, data = diam)
```

We make a scatter plot as follows:

```
> plot(price ~ carat, data = diam)
> abline(res)
```

Question 0.12. Make the residual plot through the following command:

```
> plot(res, which = 1)
```

This graphic marks points which are inluential. What are their indices? These values give the row names. They are the 8th, 10th and 14*th* row in the `diam` data set.

Question 0.13. To investigate how the regression line is influenced by these extra points, we can remove them from consideration. To do so, we can create a new data set:

```
> diam1 <- diam[-c(8, 10), ]
> res.drop <- lm(price ~ carat, data = diam1)
> abline(res.drop, col = "red")
```

Add the line above to your scatter plot (you can't close the scatter plot window or this won't work). Does it seem like the slope of the regression line has changed?

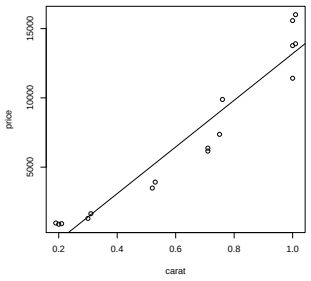

Figure 4: Scattert plot of price by carat size for "D" grade diamonds only.

Question 0.14. The Cook's distance for each point is based on looking at the change between the fitted model for all the data and the fitted model for the data without that point. If a point is influential there will be a big change. A diagnostic plot is produced by

```
> plot(res, which = 5)
```

Make this plot. What values are flagged by numbers? These are influential points