Regression III - assumptions and miscellaneous topics

If we just want to calculate a least squares line, there are no assumptions; after all, we're not doing any hypothesis testing.

> You can do silly things like fit lines to data that are obviously curved; nothing says you can't.
>
> Once you start doing hypotheses tests for regression, then you do need to deal with assumptions. There are lots in regression...

We have five assumptions we need to deal with (see also the graphs a few pages down):

1. The data must be random (you just can't get away from this one!)

2. X and Y must have a linear (straight) relationship.

   If this is not true, then *nothing* is valid. *Nothing!*

   After all, you're fitting a *line*, and if the relationship between X and Y isn't straight, this doesn't make much sense.

   We verify the relationship using *residual plots.*

3. For each level of X, the residuals are normally distributed.

   Check this using a regular *q-q* plot of the residuals.

4. The variance of the residuals is constant.

   For example, at low values of X the variance of the residuals is the same as at larger values of X.

   Check this using residual plots.

5. Each residual is independent of every other residual.

   There should be no connection/relation between any one residual and any other residual.

   Odd patterns in a scatterplot (e.g., a lot of residuals near each other, residuals forming an obvious patter, etc.) may indicate that something is wrong here.

   This can sometimes be better seen in residual plots.

That's a lot of assumptions!

> We will not learn what to do if you violate your assumptions (it would require several more weeks of lecture).

Because assumptions are so important, though, you need to be able to figure out when you are violating the assumptions.

If you are violating the assumptions then at least you know you need to consult a statistician to help you fix the problem.

Usually the problems can be fixed using techniques like rank regression or transformations, but we don't have the time to learn these.

Obviously, we need to learn about residual plots, since they are used to identify most of the problems listed above.

Residual plots act as a "magnifying lens" for a scatterplot; they let you see problems that are sometimes difficult or impossible to see on the actual scatterplot.

There are actually several different ways of making a residual plot; we'll stick with the simplest.

Your text does something else, which isn't as easy to understand, but is essentially the same thing!

So here's how to make a residual plot:

Plot the x-values on the X axis as usual, then plot the *residuals* on the Y axis.

Doing this by hand is a bit tedious as you need to calculate all the residuals.

R can do this for us automatically (as usual).

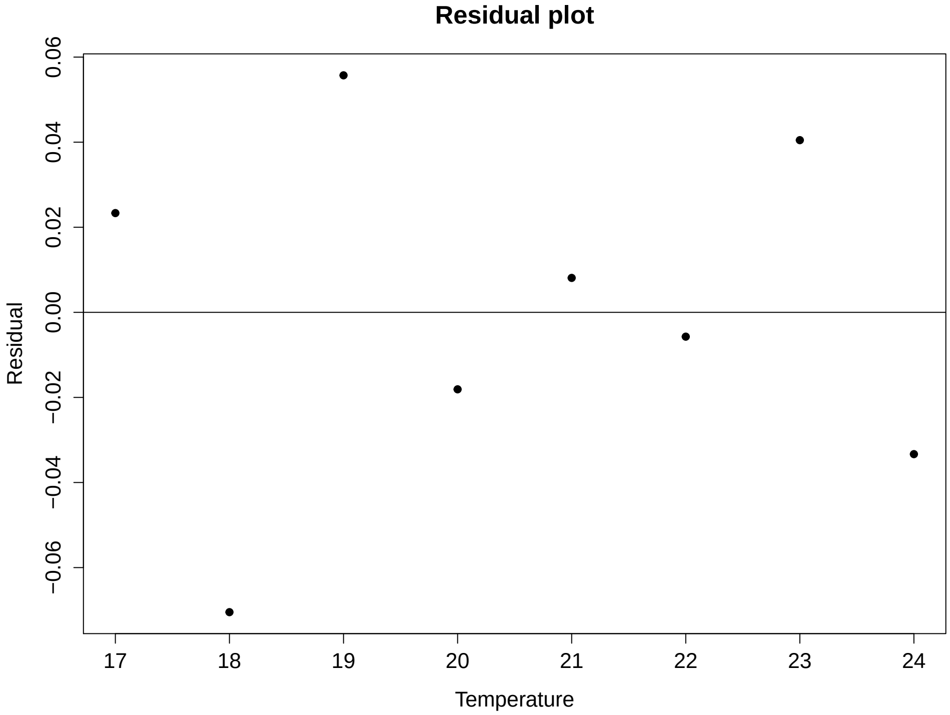
So here's an example using the previous exercise dealing with laetisaeric acid.

Here's an excerpt from the table we used when looked at hypothesis testing: (X = temperature (Celsius), Y = NCER (µmol/m^2)

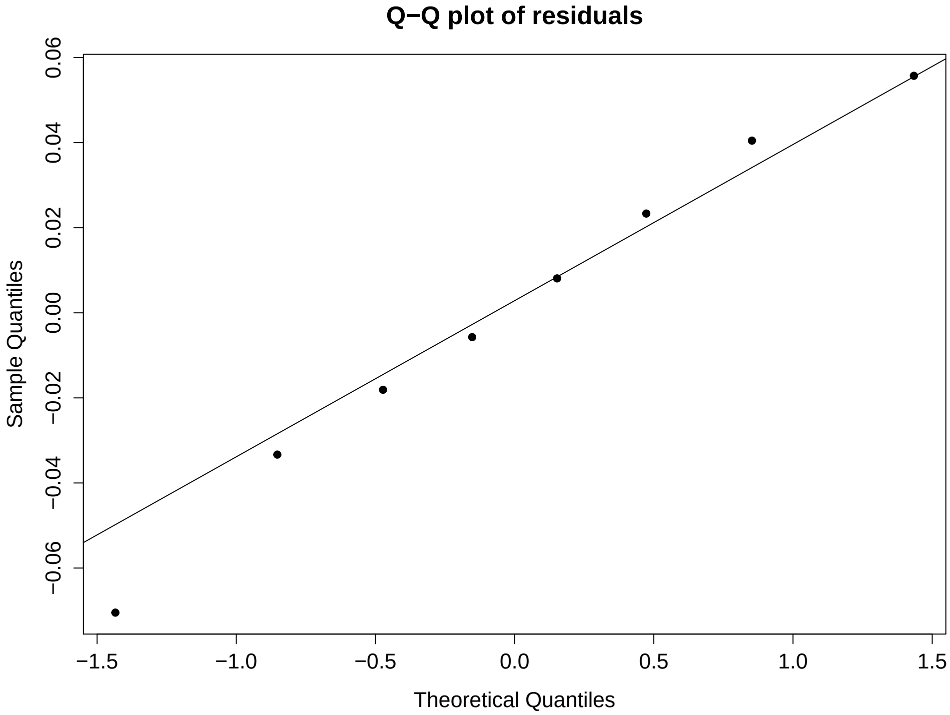
X	Y	$\hat{Y}$	r
17	0.31	0.2866667	0.023333333
18	0.27	0.3404762	-0.070476190
19	0.45	0.3942857	0.055714286
20	0.43	0.4480952	-0.018095238
21	0.51	0.5019048	0.008095238
22	0.55	0.5557143	-0.005714286
23	0.65	0.6095238	0.040476190
24	0.63	0.6633333	-0.033333333

The residuals, r, are given in the last column.

We plot x vs. r in our plot and get the following graph:



And while we're at it, we should also do the q-q plot of the residuals:



So now we know how to make a residual plot. The above plots are both pretty good plot and shows no problems with our assumptions.

So what do we look for? Or, when do we know that we have a problem? We look for:

1. An obvious curved patterns in the residuals.

   This means the straight line relationship between X and Y is violated.

2. Residuals that appear to form a funnel.

   This violates the constant variance assumptions.

3. Residuals that do both - form a curved pattern and a funnel (a "horn" type of pattern).

   This violates both the straight line relationship and constant variance.

4. Any other obvious patterns in the residuals.

   This violates the assumption of independence (it can also show other obvious problems like outliers).

The graphs on the next couple of pages show some examples of problems in residual plots.

(And don't forget to look at your q-q plot to check for the normal distribution of the residuals.)

So what do you do if you see any problems in your residual plots (or in your q-q plot)?

Talk to a statistician who can help you fix these problems.

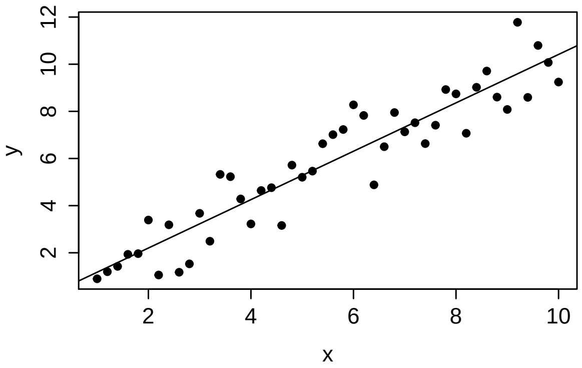
In particular, if you see a curve, your results are ***garbage*** until you straighten out the curve.

Non-constant variance generally means you loose power - you should fix it, but it's usually not quite as serious as a curve.

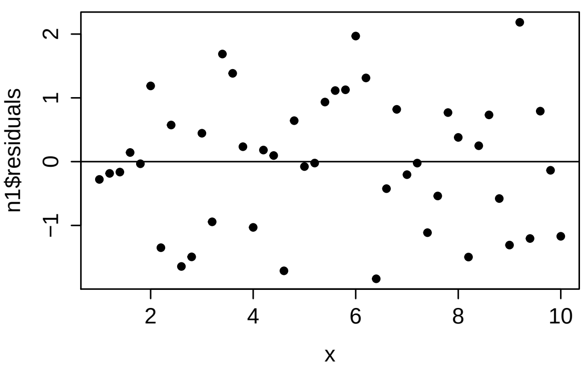
Residuals that are not normally distributed also mess up your power and should be fixed.

You might get some help from the Central Limit Theorem if you have a nice large sample size.

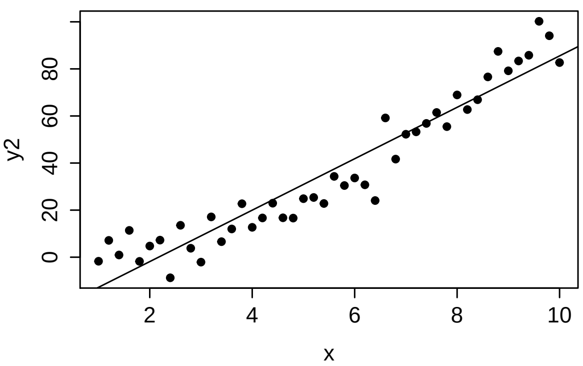
When you do a regression (or look at someone else's regression), *the residual plot should always be the first thing you look at.* Do this before looking at p-values, results, and so on!



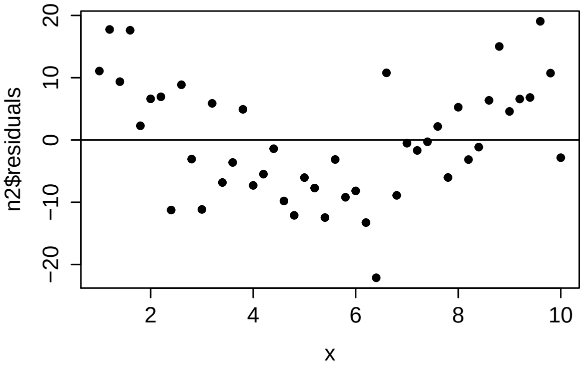
A perfectly good regression



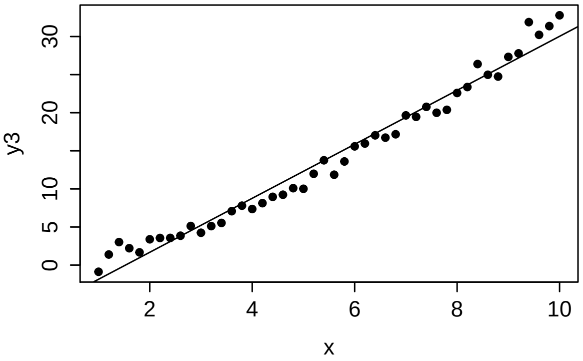
No particular patterns in the residual plot



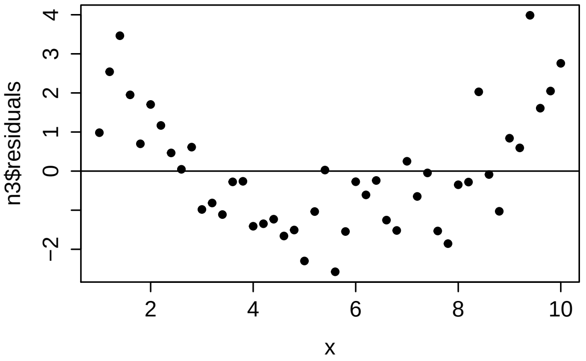
Note curve in original data



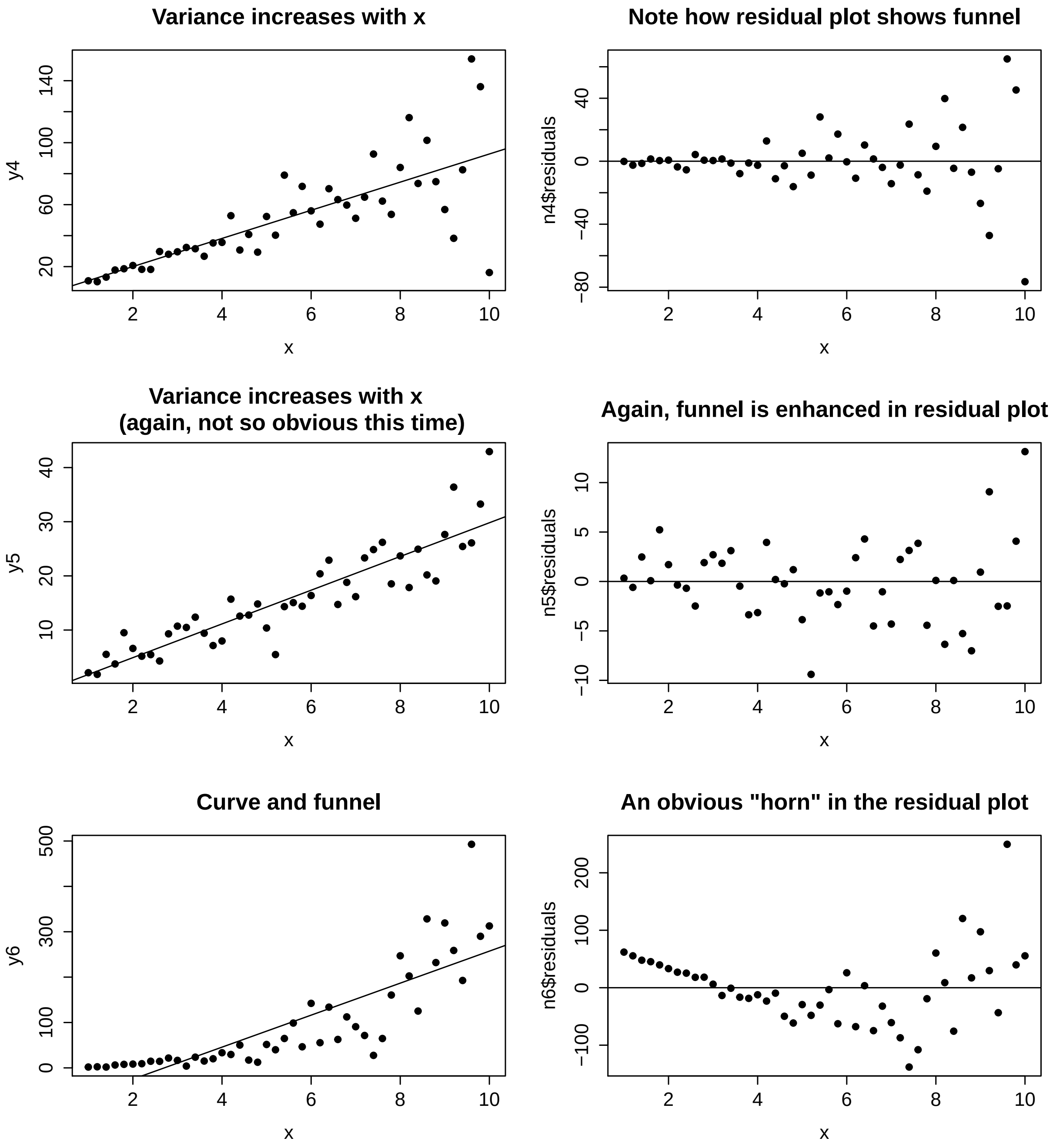
Curve is enhanced in residual plot



Note curve in original data
(perhaps not quite so obvious here)



Again, curve is enhanced in residual plot



Variance increases with x

Note how residual plot shows funnel

Variance increases with x
(again, not so obvious this time)

Again, funnel is enhanced in residual plot

Curve and funnel

An obvious "horn" in the residual plot

Other topics having to do with regression:

R^2 (not the software...)

R^2 is also known as the *Coefficient of Determination*.

R^2 measures how good our regression fits the data.

It gives us a measure to let us know how well our regression explains the relationship between x and Y.

A more precise definition is given after we figure out how to calculate it.

How to calculate R^2:

There are several ways of doing this. One is quite easy since we already know how to calculate the correlation coefficient, r:

$$R^2 = r^2$$

In other words, it's just the square of our correlation coefficient.

But this doesn't really let us understand how it works; there's another way to calculate R^2 that makes it a little easier to understand:

$$R^2 = \frac{SS_{reg}}{SS_y} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

So that looks a little messy, but it's not too bad. But before we look at it in more detail, notice two things:

$SS_{reg} \neq SS_r = SS_{residuals}$

We're using r for several different things at this point (residuals, correlation coefficient). You'll have to keep things straight.

So back to our equation. Suppose all our points are on a perfectly straight line:

This means that each $\hat{y}_i = y_i$.

Which implies that the numerator and denominator are identical.

In other words, $R^2 = 1$.

Also notice that the further away the $\hat{y}_i$'s get from the y_i's, the closer R^2 gets to 0.

So R^2 gets closer to 1 the better the relationship is between X and Y and closer to 0 the worse this relationship is.

R^2 tells us the proportion of variability in Y explained by X.

For example, if $R^2 = 0.95$, this tells us that 95% of the variation in Y can be explained by X (that would be very good).

On the other hand, if $R^2 = 0.12$, then only 12% of the variation in Y can be explained by X. This isn't very good.

This means we probably should look for another explanation for the variability in Y (even if the regression is significant).

You know how to calculate R^2. It's a useful measure of how good your regression is.

For the soil respiration problem, for example, it's pretty easy since we have everything we need already:

$$r = \frac{SS_{cp}}{\sqrt{SS_x\, SS_y}} = \frac{2.26}{\sqrt{42 \times 0.1334}} = 0.955 \implies R^2 = 0.955^2 = 0.912$$

Which is remarkably good (91.2 percent of the variation in CO_2 respiration is explained by temperature).

A significant regression doesn't necessarily tell you that you've explained everything - R^2 helps with this.

You should get in the habit of always reporting R^2 when you do a regression.

So let's summarize some of this:

When you do a regression you should:

Calculate your least squares line.

Do your hypothesis test.

Check your assumptions. *Do this before figure out if your regression is significant.*

The results of the hypothesis test (significant or not) are *irrelevant* if you're violating your assumptions (particularly the linear relationship between X and Y).

If you're assumptions are okay, go on, report the results of the test and give R^2.

Some final comments on regression:

Regression is a fairly complicated topic. Whole classes are taught on regression, and numerous textbooks are available that deal with regression.

Regression includes such topics as:

Multiple regression (using more than one X).

Polynomial regression (modeling curved data).

Multivariate regression (using more than one Y).

Non-linear regression (another way of dealing with curved data).

The interesting thing is that (believe it or not) almost all the techniques we learned in this class can be *modeled* as a regression problem.

(We can use regression techniques to do two sample tests, etc.).

A regression class is probably one of the most useful advanced statistics courses you can take - it teaches you more about practical statistics than most other courses.

You do need to brush up on your matrix algebra, though!