Linear Regression

1 Returning to Non Independent Random Variables

Much of the previous work has been focused on studying a single distribution, by observing many *independent* random variables, having that same distribution. Independence made for very convenient simplifications in our work.

However, we often have to deal with a different problem: there are two quantities, and we would like to know how they influence each other (of course, we could consider the case of more than two as well, but that increases the complications considerably, so we won't). You have seen many such cases: does smoking have something to do with getting cancer? does the stock market react to changes in the weather? <put your favorite question here>?

Unless we have some powerful theory to answer such questions (and, outside of physics, that's rarely the case), the best we can do is observe the two quantities an see if they seem to influence each other or not. As usual, bare data will tell us exactly nothing. *We need a model.*

1.1 A Mathematical Model for Dependence

We already have a model for dependence: we consider our two quantities to be modeled by two *random variables*, and we can try to figure out if knowing the value of one changes our beliefs about the other. That is, we *can try to figure out if the* ***conditional distribution*** *of one, say Y, given the other, say X, is or isn't different from the original distribution.*

As a silly example, if X is a two-valued random variable, telling us whether a die turned out or or even, and Y is a random variable telling us how many points came up on that die, we have that

$$P[Y=k]=\frac{1}{6}$$

for $k=1,2,...,6$, but

$$P[Y=1|X=\text{odd}]=\frac{1}{3},\ P[Y=1|X=\text{even}]=0$$

and so on...

On the other hand, we are probably safe if we say that the probability of throwing 1 doesn't change if the weather across the globe is rainy or not, nor does it change depending on the color of the shirt we are wearing.

Unfortunately, it takes a lot of work to determine (even only approximately) the full conditional distribution of one variable given another: we have to check all (or, at least, very many) possible combinations. While this is the sure way to go, given its difficulty, it is understandable that a less cumbersome way has been explored

1.2 Independence and Correlation

We have already mentioned that two independent random variables are also *uncorrelated*: this showed up when we discussed the variance of the sum of two such variables. In general, as you may recall, it is true that

$$\text{Var}[X\pm Y]=\text{Var}[X]+\text{Var}[Y]\pm\text{Cov}[X,Y]$$

where

$$\text{Cov}[X,Y]=E[\text{XY}]-E[X]E[Y]=E[(X-E[X])(Y-E[Y])$$

is called the *covariance* of X and Y.

In most applications, X and Y represent measurements of things, and thus their value depends on what units we are using: for example, if they represent lengths, you will have different numbers, depending on whether you are measuring lengths in inches, feet, miles, or light years. It is customary to consider a variation of the covariance that does not depend on what units you are using, called the *correlation*, defined as

$$\rho(X,Y) = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}$$

Given how covariance and correlations are related to moments, it is intuitive that we can work up an estimation method similar to the ones we have for means and variances of single random variables. Thus, estimating that a correlations has some non zero value does give some insight on the fact that the two variables are not independent. Unfortunately, without additional, serious, assumptions, the reverse is far from true: to know (or estimate) that two variables have correlation (or covariance) zero or close to zero means practically nothing. The following artificial example illustrates this:

> Suppose $E[X] = E\left[X^3\right] = 0$, and consider the two random variables X, and $Y = X^2$. Their covariance is equal to
>
> $$E[XY] - E[X]E[Y] = E\left[X^3\right] - 0 = 0$$
>
> but it is fair to say that they are anything but independent!

In other words, learning about the correlation of two random variables may tell you very little about their connections. There is a strong assumption that results in no correlation being equivalent to independence, but it is a very strong one.

1.3 Jointly Gaussian Variables

The previous discussion justifies the following strategy. Suppose we are looking at *a pair* of random variables, say X, Y. Suppose also that their *joint distribution*, that is the collection of numbers like

$$P[a \leqslant X \leqslant b, c \leqslant Y \leqslant d]$$

is a *two-dimensional Gaussian distribution*. Then, the variables are said to be *jointly Gaussian*. This means, not only that both X, and Y are Gaussian, but also that statements like the one above, involving both variables, are evaluated using a special density depending on two variables,which is called a 2-dimensional Gaussian.

Note 1. The precise mathematical formulation is not necessary here, but here it is. A one-dimensional Gaussian distribution is such that $P[a \leqslant X \leqslant b]$ is equal to the area below the curve $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ and the x axis, between the points a and b. A *two-dimensional Gaussian* distribution is such that the probability above is equal to the *volume* below the surface, in 3-dimensional space, that is the graph of a function like $\frac{1}{2\pi K} e^{-\frac{1}{2}\left(A(x-\mu)^2 + B(x-\mu)(y-\nu) + C(y-\nu)^2\right)}$ (where K is an appropriate constant, so that the total volume under the surface is equal to one), over the rectangle $a \leqslant x \leqslant b, c \leqslant y \leqslant d$. All we are saying in this section is generally false if the two variables are Gaussian, but not jointly Gaussian.

Under this assumption (and a similar statement holds for a "jointly binomial" case). we have the remarkable result:

Theorem 2. *If X and Y are jointly Gaussian and uncorrelated (their covariance, and, hence, their correlation, is equal to zero), they are independent*

Please, remember that this can be completely false without the jointly Gaussian assumption. Whether the assumption holds in any given situation may not be easy to determine, and, as it often happens, you might find cases where people simply shortcut the issue, and zoom on checking the correlation of two variables without bothering too much with the niceties of this theorem.

That said, there is more to the story, as long as the jointly Gaussian assumption is justified. In fact, in this case, we can actually nail down the full conditional distribution, once we have determined the covariance/correlation of the two variables! To see how this can work, and how this can provide us with an amazingly powerful forecasting tool, we have to move further, and try to provide a complete model of dependent Gaussian variables (but, keep in mind that much of the construction is contingent on the Gaussian assumption).

I Gaussian Case: A Rigorous Case For Linearity

2 Why A Model for Dependent Variables?

We have two random variables, that we expect to be dependent, and we would like to express their dependence explicitly. That is what we do when we do regression analysis. But why would we bother?

The main point of constructing a solid model for how a variable, say Y, depends on another, say X, is (as in most scientific endeavors) **so we can predict what value of Y will be observed, if X takes on a specific value**. Thus, if we had (we don't, but it would be great) a good model for how the unemployment rate depended on the discount rate set by the Federal Reserve, we would have a powerful tool to solve the unemployment problem. Of course, we would not expect such a dependence to be mechanical, but rather *probabilistic*: it would be "very likely" that unemployment would fall within a certain interval, if the discount rate was set at, say, 2.5%.

While this simple dependence between interest rates and unemployment is well known not to exist in the real world, there are less momentous pairs of quantities where one can extract reasonable predictions once a good dependency model has been found. That is the role of regression analysis: **predict what has not been observed, on the basis of past observations**.

The surprisingly simple tool for this is what is called the "Least Mean Squares" method. In brief, you look at X and Y as a pair, observe them *together*, so that you can get a hold on the distribution *of the pair*, and you try to approximate the "cloud" of data with the "best" straight line you can come up with. "Best" is defined, in this context, as *the line that corresponds to the least mean square error*.

2.1 Theoretical Least Mean Squares

Under the assumption of joint Gaussian distribution, it can proved that we can always write

$$Y = aX + b + \varepsilon$$

where a, b are real numbers, and ε is a Gaussian random variable, whose mean is zero, and is independent of X. What is more, a and b can be determined by choosing them so as to make the quantity

$$E\Big[\,(Y - aX - b)^2\,\Big] = E\big[\,\varepsilon^2\,\big]$$

as small as possible. Of course, to do this, we need to know the true distribution of the pair X, Y, which we normally will not.

Remark 3. We can make this more specific. Under the assumptions we made, **the expression $aX + b$ can be interpreted as the expected value of Y, conditioned on the value of X** (that is, we consider the distribution of Y, when we condition it on a specific value of X—they are not independent in this model—and take the corresponding expected value. The number $E\big[\,\varepsilon^2\,\big]$ happens to be the corresponding *conditional variance*, and, in conclusion, the whole procedure determines the distribution of Y, if we assume to know the value of X (it so happens that this is a Gaussian distribution, and hence identifiable once we know its mean and variance). Given the distribution of X, we can then reconstruct the distribution of Y, and, in fact, the joint distribution of X, and Y.

Remark 4. The whole procedure, as you my not see immediately, **is asymmetric: the roles of X and Y are distinct!** In particular, if we switch their roles, and try to provide a least men squares estimate of $X = a'Y + b' + \varepsilon'$, the numbers we get **are not what we would get by solving the previous LMS estimate for X!**

Note that if $Y = aX + b + \varepsilon$, $E[Y] = aE[X]$, and $E[XY] - E[X]E[Y] = aE\big[X^2\big] + bE[X] - a(E[X])^2$ (recall that $E[\varepsilon] = 0$, and $E[X\varepsilon] = E[X]E[\varepsilon] = 0$, since X and ε are independent), that is

$$\mathrm{Cov}[X, Y] = a\,\mathrm{Var}[X] - bE[X]$$

2.2 The Least Mean Squares Formulas

The problem of finding the values of a and b that will minimize the "mean square error" is not difficult to solve, but you do need some tools. Most books will tell you that you need some calculus, and this certainly works, but you can get away with less: a little geometry of ellipses will do the work.

The result is

$$a = \frac{\mathrm{Cov}[X, Y]}{\mathrm{Var}[X]}, b = E[Y] - aE[X]$$

If you are curious, the proof is as follows (this is obviously for your information only)

2.2.1 Proof of the Least Mean Square Formula

Suppose we were able to compute all expectations explicitly, having a full description of the joint distribution of our pair (X, Y). Then, we could, for any choice of a, b, compute

$$\begin{aligned} E\Big[(Y - aX - b)^2\Big] &= E\big[Y^2 + a^2X^2 + b^2 - 2aXY - 2bY + 2abX\big] = \\ &= a^2E\big[X^2\big] + b^2 + 2abE[X] - 2aE[XY] - 2bE[Y] + E\big[Y^2\big] \end{aligned}$$

If we look at this expression as a function of a and b, it describes and ellipse in a hypothetical $a - b$ plane every time we fix a value for the whole expression. It is easy to see that the higher the value, the larger the ellipse, and that all these ellipses are all concentric. Hence, the choice of (a, b) that makes this expression the smallest is the common *center* of all these ellipses. This is easy to determine, if you have studied the geometry of ellipses that may be oriented obliquely in the plane. Even if you haven't we can find the center with a couple of tricks.

The simple case: $b = 0$ Suppose that, for some reason, we think that we expect $X = 0$ to always correspond to $Y = 0$ (it could be because of the specific application, one cannot have a non-zero value if the other is zero). Then the expression we found is simpler:

$$a^2E\big[X^2\big] - 2aE[XY] + E\big[Y^2\big]$$

This, as a function of a, is a parabola, concave up, and has a minimum at its vertex, which is at (from your algebra classes)

$$a = \frac{E[XY]}{E[X^2]}$$

Hence, the least mean square zero-intercept line is given by

$$Y = \frac{E[XY]}{E[X^2]}X$$

The case $E[X] = 0$: If X has zero expectation, the expression is

$$a^2E\big[X^2\big] + b^2 - 2aE[XY] - 2bE[Y] + Y^2$$

To find the center of this ellipse, we have only to "complete the square", as you have already seen in your precalculus classes. Note that we don't care for the full operation: we only need the center, so we only need to find what is missing to complete the two squares in a and b. For a we have

$$a^2E\big[X^2\big] - 2aE[XY] = E\big[X^2\big]\left(a^2 - 2a\frac{E[XY]}{E[X^2]}\right)$$

and it is now clear that the center will correspond to $a = \frac{E[XY]}{E[X^2]}$, as in the simple case above. The value of b is even faster to find:

$$b^2 - 2bE[Y]$$

shows that the center is at $b = E[Y]$.

The general case:

If you have taken a Calculus class, you know how to find the minimum of our expression in general: you compute its derivative, considering it only as a function of a, and set it equal to zero, you then compute its derivative considering it only as a function of b, and set this to zero. The two equations in a and b that you get are linear and easily solved:

$$\begin{aligned} 2aE[X^2] + 2bE[X] - 2E[XY] &= 0 \\ 2b + 2aE[X] - 2E[Y] &= 0 \\ b = E[Y] - aE[X] & \\ aE[X^2] + E[X]E[Y] - a(E[X])^2 - E[XY] &= 0 \\ a = \frac{E[XY] - E[X]E[Y]}{E[X^2] - (E[X])^2} & \end{aligned}$$

Notice how the solution for a is the ratio of the covariance of X and Y, divided by the variance of X.

You can get the same result, with a little work, using only algebra, though. For this purpose, we write our problem not for the variable X, but for the variable $\tilde{X} = X - E[X]$. Of course, $E\left[\tilde{X}\right] = 0$, so we are in the previous case, and find, for the regression $Y = \tilde{a}\tilde{X} + \tilde{b}$

$$\tilde{a} = \frac{E\left[\tilde{X}Y\right]}{E\left[\tilde{X}^2\right]}, \tilde{b} = E[Y]$$

that is

$$\begin{aligned} Y &= \frac{E[(X - E[X])Y]}{E\left[(X - E[X])^2\right]}(X - E[X]) + E[Y] \\ Y &= \frac{E[XY] - E[X]E[Y]}{\mathrm{Var}[X]}X + E[Y] - E[X]\frac{E[XY] - E[X]E[Y]}{\mathrm{Var}[X]} \end{aligned}$$

that is, back to our original a and b, in $Y = aX + b$,

$$a = \frac{\mathrm{Cov}[X, Y]}{\mathrm{Var}[X]}, b = E[Y] - aE[X]$$

which is the same result we find if we use calculus, as well as matching our proposition 5.

We can also write

$$a = \rho(X, Y)\sqrt{\frac{\mathrm{Var}[Y]}{\mathrm{Var}[X]}}$$

As a side result, we may notice that $\mathrm{Var}[Y] = \mathrm{Var}[aX + b + \varepsilon] = a^2\mathrm{Var}[X] + \mathrm{Var}[\varepsilon]$ (you may want to check this in the probability module), so that

$$\rho[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} = \frac{a\,\mathrm{Var}[X]}{\sqrt{a^2(\mathrm{Var}[X])^2 + \mathrm{Var}[X]\mathrm{Var}[\varepsilon]}}$$

equivalent to $\rho(X, Y) = a\sqrt{\frac{\mathrm{Var}[X]}{\mathrm{Var}[Y]}}$, but with no explicit use of data from Y.

3 Application To A Sample

Of course, in the situations we are interested in, we do not have the joint distribution, so the previous discussion will seem completely theoretical. But here is the way the problem is solved in statistics:

> To find our "best" estimate for the coefficients a and b, we will do the same calculations – and hence arrive at the same formulas – **using the empirical distribution** of our sample, in place of the (unknown) "true" distribution.

If you go and look what the formulas for *estimating* the "best fit" line and/or the correlation between two variables are, you will notice that they look strikingly similar to the formulas in the previous section, once you make a crucial connection.

Recall that the empirical distribution of a sample x_1, x_2, ..., x_n is the probability distribution that assigns probability $\frac{1}{n}$ to each value x_k. If we go and look at formulas that estimate the least mean square approximation of one variable to another, the formulas are exactly the ones in the previous section, **provided we use he empirical distribution in place of the "true" one**.

Thus, for example, the expected values are replaces by the empirical means, and the covariance and variance by their empirical analogs, $\frac{1}{n}\sum_{k=1}^{n} x_k y_k$, and $\frac{1}{n}\sum_{k=1}^{n} x_k^2$. In published formulas, the factors n are rearranged for looks, so it might not be immediately obvious that we are dealing with empirical moments as building blocks for the solution.

3.1 Review: Theoretical and Empirical Distributions

You may recall from a previous module the following scheme, which is at the foundation of statistical reasoning. The scheme goes as follows.

1. We assume that an observation can be represented as observing a random variable X. Repeated observations, if performed properly, should correspond to successive observations of random variables $X_1, X_2, ..., X_n$, independent, and all with the same distribution as X.

2. X has a distribution which determines things like $P[a \leqslant X \leqslant b]$, as well as parameters like $E[X]$, $\mathrm{Var}[X]$, and so on. We are trying to determine this distribution, or, at least, some of its parameters.

3. What we have are our observations: $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$. These represent a random occurrence: if we should repeat our experiment, we would get different values. Now, we can think of these observations as constituting a *random distribution* (called the *empirical distribution*), by assigning weight $\frac{1}{n}$ to each observed value. It turns out (it is an extension of the Law of Large Numbers) that, under reasonable assumptions, if we consider larger and larger samples, as n grows, the number of observations between a and b will get closer and closer to $P[a \leqslant X \leqslant b]$, and, of course, $\frac{1}{n}\sum_{k=1}^{n} x_k$ (the "expected value" of the empirical distribution) will get closer and closer to $E[X]$ (that's exactly the LLN), and so on.

Now, the facts above suggest that we try to use the empirical distribution in place of the theoretical one (which we don't know). It won't be exact, but, for large samples, it won't be too bad. Also, the statements above can be refined, so we can actually give some precise meaning to "close".

You can now see what is going on in the Linear Regression model: assuming that the pair of variables we are working with are jointly Gaussian, our procedure consists in substituting their *empirical distribution* for their theoretical distribution in the search for a and b as discussed here. Theorems in the spirit of the LLN will guarantee that our (empirical) coefficients will get closer to the "true" ones for large samples, and theorems in the spirit of the CLT will allow us to produce the analog of interval estimates, as well as tests, much like we did in the previous modules.

Note 5. The expression we are minimizing becomes, in the sample case,

$$\frac{1}{n}\sum_{k=1}^{n}(y_k - a x_k - b)^2 = \frac{1}{n}\sum_{k=1}^{n} y_k^2 + \frac{a^2}{n}\sum_{k=1}^{n} x_k^2 + b^2 - \frac{2a}{n}\sum_{k=1}^{n} x_k y_k - \frac{2b}{n}\sum_{k=1}^{n} y_k + \frac{2ab}{n}\sum_{k=1}^{n} x_k$$

This is often called "the sum of the residuals", and, with a little manipulation to get rid of the amplifying effects due to your choice of units, can be used as a qualitative measure of how tight our approximating line will be to the real data.

II Non Gaussian Case: LMS As an Intuitive Tool

4 Why Least Mean Squares?

Regardless of any assumption of normality (let alone, of *joint* normality), you will find extensive use of least mean square interpolation of data points. Why are we using the *Least Mean Squares* (LMS) method to approximate a cloud of points with a line? An why a line, in the first place?

There are easy answers to both, but they are not as strong as the logic we discussed in the previous part.

- We look for a line because lines are easy to handle and to understand. They also allow for easy pointing out of trends
- We use the LMS method, because the math is much simpler than other options

Both items are true, but their strength is a bit the strength of convenience, rather than the strength of a solid logical foundation.

5 Leaving the Gaussian Cover

Remark 6. As already stated, the powerful theory we sketched in the previous part based on a very specific assumption about the distribution of our pair. All the indexes that any regression program will produce have a direct interpretation in this case. If this assumption cannot be made, or is outright inappropriate, the theory disappears, and we are, indeed, only left with the somewhat lame justification listed at the very beginning of this part. Note that, like most anything we do in science, that does not mean that, in itself, it is *wrong* to use LMS outside of the Gaussian framework. It would be wrong, if we pretended that our linear model meant as much as it would if the distribution was Gaussian. It is OK to use it, as long as we realize that we cannot make very strong deductions from our calculations, and that, generally speaking, we are not describing our pair of variables in the most precise and illuminating way. In the last section we will discuss what, if anything, we should make of a least mean square estimate when the Gaussian assumption is not acceptable.

The theory we have briefly sketched is mathematically sound, and provides a rigorous underpinning to the use of linear regression, provided its assumptions are a reasonable model for the practical situation we are studying. There are several directions that can be taken if we leave the comfort of the Gaussian environment.

5.1 Non Gaussian Least Mean Squares

Even if we have to drop the Gaussian assumption, we can still make sense of a least mean squares approach to estimation—the only problem is that most (or all) of the probabilistic support is lost, and we cannot use things like interval estimates and tests with much confidence, if at all.

In fact, the whole method can be cast in a non probabilistic frame, with the advantage of not requiring stringent assumptions, and with the downside that what we get is only what we see—for example, no deep information on how likely it is that our estimates will be effective for forecasting (sure, the sum of residuals will tell us how good the regression line fits *the data we already have*, but it gives no argument as to why it should be good for data that we will look for in the future). There is a quote of Laplace, arguing that minimizing the square error is, so to speak, the "natural" thing to do, but, with all due respect, the biggest advantage of this approach is its mathematical simplicity, more than its "naturality".

In essence, the method assumes that we can say that the following model is reasonable: "The two quantities X and Y are linearly dependent, up to a random, mean zero error". Minimizing the square error means then minimizing the variance of the error term. Note that, read like this, the model $Y = aX + b + \varepsilon$ often considers X *not to be random*, hence, the notion of *correlation* loses its real meaning. Incidentally, if we assume that, instead, X is in effect a Gaussian random variable, and ε is a mean zero Gaussian random variable, independent of X, then we are fully in the situation described in the first part.

The "naive" justifications listed at the very beginning of this part are not unreasonable. In particular, minimizing the square of the discrepancies between our data and our model is

- convenient: it makes for very tractable mathematics, as opposed to, say, minimizing the sum of the absolute values of the discrepancies
- more or less reasonable: it treats small deviations as not too important, but takes big deviations seriously—however, since the threshold is the number 1 ($a < 1 \Rightarrow a^2 < |a|, a > 1 \Rightarrow a^2 > |a|$), "small" and "large" in this argument depend on the units chosen.

This has downsides, of course. For example, this method is very sensitive to outliers. More damning, though, is the apparent positive feature that, **no matter what your data is, there is always a least mean square linear estimate**. That is, even if the data has no reasonable way of being interpreted as "linear dependency+error", it will still provide an answer! To be a bit more precise, in a way, you can always argue for some such model, except that, provided you don't rely on just a few data points, the "error" term will prove to be huge, and will grow, as you add data points (in principle, if the Gaussian model is applicable, adding data points should not increase the least mean square error too much—this argument can be made very precise, but we can leave it at that here, even though details are available on request).

To drive this point home, if you will go on the Internet and download unemployment data for the last N years (depending on your storage, and how far back the data is available), and Fed discount rates for the same period of time, **you will find a linear regression line between the two**. Whether this line is of any use in forecasting anything, is a different question.

5.2 Non Linear Least Mean Squares

Of course, if the data seems at odds with a "linear+error" model, let alone a jointly Gaussian assumption, we can always try something else, as in

- Polynomial (quadratic, cubic, ...) + error
- Exponential or logarithmic + error
- Some other functional dependency + error

This is indeed sometimes done. Of course, probability and statistics are almost completely out of the picture now, so tools like confidence intervals or tests are unavailable. Also, all these models are workable, essentially, only if they have "free" parameters (the analog of a and b in the linear model) that appear *linearly*. For example, a quadratic model like

$$Y = aX^2 + bX + c + \varepsilon$$

can be treated, essentially, in the same way (if you don't have calculus, you need geometry of surfaces in 3 dimensions to do the work, but you get the idea). Similarly, a model like

$$Y = ae^X + b + \varepsilon$$

is not any different from a linear model. Indeed, you only have to introduce a new variable, say, $Z = e^X$, and, the model is now linear in Y and Z. Z could not be Gaussian in this case, since it is necessarily positive, but if it could be reasonably approximated as Gaussian (as in $E[Z] = 100, \mathrm{Var}[Z] = 4$—the probability of a negative value of Z is so small, that we might as well ignore it), independent of ε, we could even bring in the whole theory from the first part.

However, a model like

$$Y = ae^{bX} + c + \varepsilon$$

becomes way more elaborate, when you try to find a, b, c that minimize

$$\frac{1}{n}\sum_{k=1}^{n}\left(y_k - a\,e^{b x_k} - c\right)^2$$

for the n data points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) (you recalled that this is what we had to do, right?). There is no way around calculus to find this minimum, and the equations we end up are not linear in a, b, c, which makes them very awkward to solve.

We might have some theoretical argument to argue for a specific "function type + error" model, but, oftentimes, we don't: it's only that "the data looks like that". When that's the case, we don't really have a solid argument (as we have under the jointly Gaussian assumption) to support any of these models, and even the criterion "choose the functional dependency that minimizes the least mean square error" does not work: if you have n data points, there always is a polynomial of degree $n-1$ whose graph goes through these n pairs exactly. It is not an interesting model, since chances are that if we make an $n+1$-th observation, it will not sit on the graph at all, but we are also without a strong theoretical argument for anything else. At best, we can argue that whatever choice we made, we tried the simplest functional form that, somehow, seemed to capture the looks of the data, and cross our fingers that this will yield useful predictions.

6 Pitfalls

If you are applying a least mean square method to some data cloud, for example to find the "best fit" straight line, there is an important fact you have to bear in mind, as we discussed above: *regardless of the structure of the data, there is always a unique least mean square line*. In other words, the procedure is "blind" and will spit out a "best fit" line, regardless of whether a linear model makes sense or not.

Of course, as you may explore also referring to the more extensive discussion in the On Line Stat Book, the "residuals" (the sum of squares that we have minimized) will be huge, if a line is not a good approximation. Nonetheless, if you are not looking, you can come up with pretty nonsensical results.

In a way, whenever a joint Gaussian model is not really justified, linear (or, for that matter, nonlinear) regression is more part of *descriptive statistics*, than of inferential statistics—even though, since we have things like variances, and errors, this not an accurate statement. This should be meant in the sense that, in such generic cases, whatever regression method you use has to be thought of as a tool to summarize (in a way, arbitrarily) your data, rather than providing a model for the data, in the strict sense of the word. This is a delicate subject: since your choice of a straight line, or some other curve, is, at the end, based on your intuition, you are not working on an objective basis, but rather providing some numerical underpinning to your *feeling* for the data. This is a perfectly legitimate operation, but it is quite different from one that presents itself as a full-fledged objectively based prediction model.