Models In Epidemiology And Biostatistics
Gordon Hilton Fick

The Assumptions For Linear Regression

Assumptions:

1. $E(y)=\sum_{j=0}^{k}\beta_j x_j$ The conditional mean is a linear combination of the explanatory variables
2. $Var(y_i)=\sigma^2$ The conditional variance is constant.
3. $y_1\ y_2\ y_3\ \ldots\ y_n$ are statistically independent.
4. The conditional distributions are normal distributions [This assumption makes the inferences exact] or, at least, the conditional distributions are symmetrical [to provide meaning for the conditional mean]

If all assumptions, but #1, hold, it may be possible to use a nonlinear model with least squares. By nonlinear, we mean that there no way to express the conditional mean as $E(y)=\sum_{j=0}^{k}\beta_j x_j$. Notice that a model like $E(y)=\beta_0+\beta_1 x+\beta_2 x^2$ is still called a linear model in so far as the conditional mean is a linear combination of the explanatory variables [1, x and x^2]. But models like:

$E(y)=\beta_0+\beta_1\beta_2^{x}$ [exponential]

or like: $E(y)=\beta_0+\dfrac{\beta_1}{1+e^{\beta_2(x+\beta_3)}}$ [logistic]

are nonlinear.
We would still call a model for a micro assay studying the relative potency k [of the test relative to the standard] like: $E(y)=\beta_0+\beta_1 x+\beta_1(k-1)dx$ a linear model because if we were to write $\beta_2=\beta_1(k-1)$ then we get $E(y)=\beta_0+\beta_1 x+\beta_2 dx$ [d is the indicator for the test version of the drug]

Assumption 1 is a bit slippery. The conditional mean must be linear in the explanatory variables; the right set of them though. So, for example, if the conditional mean is, in fact, a quadratic in x and $E(y)=\beta_0+\beta_1 x+\beta_2 x^2$ [i.e. linear in 1 x x^2] and the model under consideration is only $E(y)=\beta_0+\beta_1 x$, then assumption 1 is not held by the model under consideration. Similarly, the absence of a necessary product of terms invalidates assumption 1. Models capturing multiplicative effects may not be expressible according to assumption 1 as well.

If all assumptions, but #2, hold, it may be possible to model the form of the conditional variance and then use a technique called 'weighted' least squares. The weights are then chosen to be proportional to the reciprocals of the variances. [Gauss-Markov].

If all assumptions, but #3, hold, it may be possible to adjust the methods if the form of the lack of independence is known. In particular, there are methods that handle matching or blocking or clustering. There are methods that are specific for longitudinal studies. Direct empirical assessment of independence is typically difficult. Most times, researchers must be aware that their study design may require one of the more specific methods. The independence assumption cannot be overlooked. Studies that involve appropriate elements of randomness [like simple random sampling] and/or randomization and for which an outcome is measured only once per subject usually enable the assumption of

independence.

If all assumptions, but #4, hold, there are a vast list of options available to the analyst. Often, if assumption 4 is not tenable, then other assumptions are questionable as well. For example, if the conditional distribution is skewed, then the conditional variance is most likely not constant. As noted earlier, tests and confidence intervals may still have their [approximate] sampling properties even if the conditional distributions are skewed so long as the sample sizes are 'large enough'. Nevertheless, the absence of symmetry may make the interpretation of the conditional means somewhat shallow and, potentially, of little scientific value.

Now we will add to the mix, the challenge of a sample that contains one or more observations that plain and simply do not follow the model that the others do follow. The reasons for such violations are many. Maybe inclusion/exclusion criteria were not followed for one or more subjects. Maybe the measuring device fails in some way occasionally. Maybe data entry has imperfections. The data entry clerk cannot have perfect methods. Errors/blunders/failures can and do happen.

“The occurrence of observations we do not like is the commonest feature of all experimental and other statistical inquiries.” (D.B. DeLury)

An implicit assumption [maybe a fifth assumption] is that our data does not contain any of these violations.

Analysts will use methods designed to find these violations. Inevitably, these methods, can, on occasion, find 'so-called' outliers. The detection of outliers can be important in so far as the investigator may then know that there are violations. Outliers are detectable violations sometimes. Outliers, though, can be truly legal [in the sense that the model still holds] but they are unusual. And.... so... what we do with the outliers?

“The decision to reject observations should never be reached lightly. The decision to reject is a decision that the error system is out of control and we lose the essential basis for reaching assured conclusions. In a way, the concern is less about the observations we remove than the ones we retain. How trustworthy are they if the error system is not to be trusted? “ (D.B. DeLury)

Outlying observations need to be checked, if at all possible, using the original raw records. Perhaps the reason for outlyingness can be determined. Of course, simply deleting an observation because we do not like it is not the basis for strong science. If the review process reveals trouble in the study, then, at best, the investigator needs to describe such trouble under limitations/qualifications in any manuscript. There are automated methods for outlier deletion out there, this writer cannot recommend such automated methods. Their use is a strong indication that the study is in big trouble. How can generalizability be possible is such muddled messes?

“...it would still be true that the Natural Sciences can only be successfully conducted by responsible and independent thinkers applying their minds and their imaginations to the detailed interpretation of verifiable observations. The idea that this responsibility can be delegated to a vast computer programmed with Decision Functions belongs to the fantasy of circles rather remote from scientific research.” (R.A. Fisher)

Assessment of Assumptions and Data Transformations

It is generally acknowledged that any one study of a particular outcome is not going to provide the

basis for an assessment process. Often, health research is not conducted with large enough sample sizes to enable empirical assumption assessment in some definitive way. Further, the soundest methods for assumption assessment are of largely graphical nature. It is probably unfair to expect that health researchers will be able to stare at a graph and use this staring to make some assessment. Graphical assessment is not easily learned. There are endless shades of gray. It is rare to see a display that resolves one or more of these matters. Having said all this, there are definitely some graphical methods that are superior to other graphical methods. There are also some misconceptions about certain strategies for graphical assessment.

Even before beginning an empirical assessment, it is best to review the literature for currently accepted assumptions and to consider well established strategies to enable reasonable assumption validity.

For example, it is generally accepted that skewness in a distribution is inevitable in variables capturing duration [of illness] , length [of hospital stay], distance [from a smelter] and other similar variables.

Another example is variables based on ratios [such as BMI]. These distributions are most often skewed as well.

The commonest resolution to this form of skewness is to use a logarithmic transformation. Sometimes, the investigator will consider a transformation from x to y like y=log(x+A) where A is chosen to correctly handle data values of x=0.[if x=0 is possible]

It is worth repeating that symmetry (or absence of skewness) is valuable for the conditional distribution of the response. Skewness in an explanatory variable is not, per se, a concern since all statements from regression analyses are conditional statements given a particular set of values for the explanatory variables. At the stage of interpretation, the possible distribution of the explanatory variables is not relevant. On the other hand, there are many circumstances in which the linearity with an explanatory variable is enhanced by a transformation that also, coincidentally, improves symmetry.

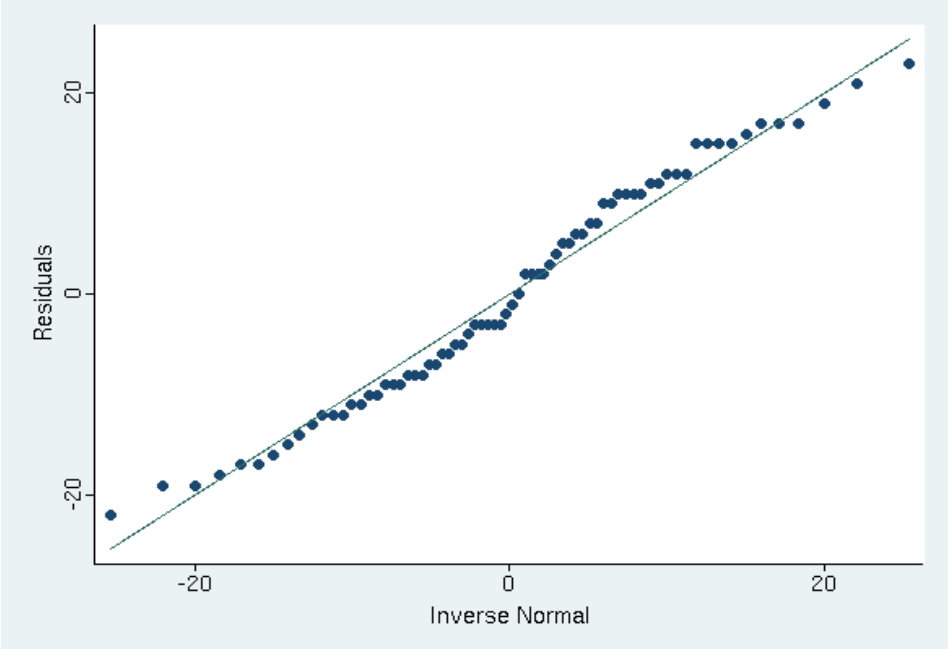
If a transformation is planned for an outcome, then interpretation may be enhanced if corresponding explanatory variables are transformed as well. For example, consider a study of household water consumption in the year 2005. It may be relevant to consider household water consumption in, say, the year 2000 as a potential explanatory variable. If household water consumption is transformed say using logarithms, then both variables [2000 and 2005] ought be transformed.

The assumptions about distribution form are based on the conditional distributions. If there are just a small number conditions under consideration, then you may be able to plot graphs of the conditional distributions for every distinct combination of conditions. We did this with the blood pressure example by considering the boxplot of blood pressure for the 12 different groups. You may be tempted to graph the distribution of blood pressure ignoring the conditions. Unfortunately, this graph tells us little about the individual conditional distributions [unless all 12 groups have the same conditional mean] since this single combo “distribution” may be distorted by the possibly different conditional means. A set of skewed conditional distributions can look symmetrical when combined in a single plot. Similarly, a set of symmetrical conditional distributions can look skewed when viewed in a single plot.

One needs to strip off the conditional means from the individual conditional distributions if one wants to try to assess distribution form on the basis of a single graph. Here is where the residuals help. The residuals e_i are estimates of the errors ϵ_i . If all other assumptions are holding reasonably, then the residuals should provide for an assessment of distribution form of the errors. A simple boxplot of the residuals can help to detect outliers. A more sensitive display is called the q-q plot. This has the

residuals on the vertical axis and the quantiles of the standard normal on the horizontal axis. This graph will show a roughly straight line if normality is plausible. Right skewed residuals will veer above the line on the right hand side while left skewed residuals will veer below the line on the left hand side.

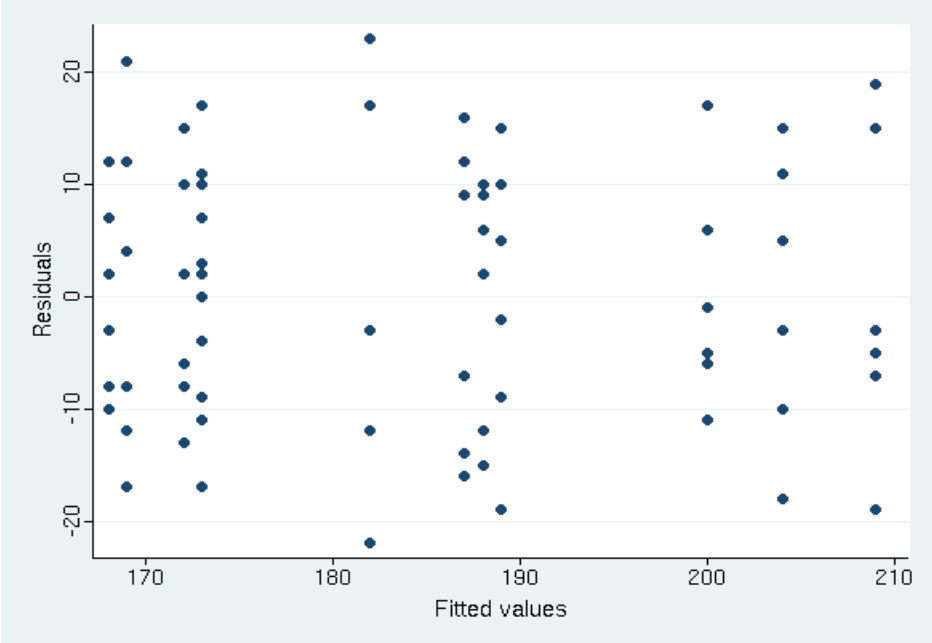
```
regr bp Y Z S B YS ZS YB ZB SB YSB ZSB
predict bph
predict res,residuals
qnorm res
```



The blood pressure residuals show no sign of skewness. No visual evidence against the symmetry assumption.

Graphing the residuals versus the fitted values can be helpful in the assessment of the constant variance assumption. No visual evidence against the constant variance assumption.

```
scatter res bph
```

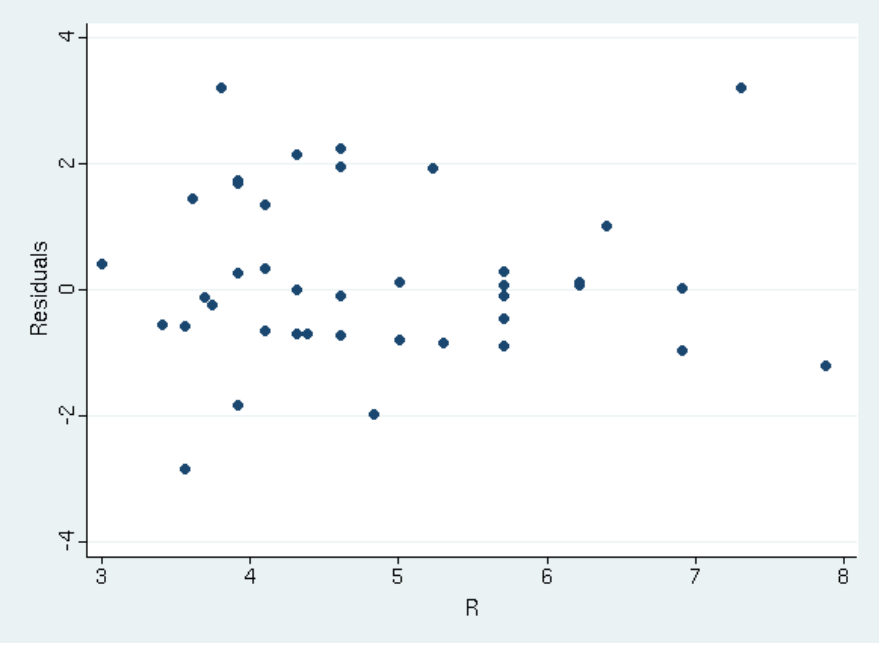


Here, there is no indication that the variability of residuals is somehow dependent on the conditions, or the conditional means. Systolic and diastolic blood pressure are 'well known' to be symmetrically distributed. So most researchers would probably note that these graphs are consistent with the commonly accepted assumption. Such displays are usually constructed as a check though. Such displays can reveal outliers in a more explicit way than other methods. The presence of such outliers can distort the fitted values and damage the value of the residuals as well. If an outlier is influential, most analysts would say they are in trouble. Tough road ahead, indeed. There are those who would argue that detection of such trouble is not worth the trouble. Something like, what we do not know, cannot hurt us. Ostriches might agree with this view.

There are numerous possible enhancements to such visual assessments. Some authors suggest plotting the absolute value of the residuals versus the fitted values. Some suggest adding a smoothed version of the absolute value of the residuals [lowess smoothers are highly regarded, these days]

There are many options available in the assessment of the assumption of linearity of the response to specific explanatory variables. Generalized Additive Models (GAMs) [gam in Stata but MSWindows only], Restricted Cubic Splines [mkspline in Stata] and Fractional Polynomials [fp in Stata] are often used. In the context of linear regression, we can graph the residuals versus individual specific explanatory variables as another potentially valuable way to assess linearity. Have a look at the wells example again.

```
regr lc R W WR
predict res,residuals
scatter res R
```



If this graph showed any 'curvature' [if we could 'see' a U shape or a parabolic shape] we might have visual evidence against linearity of log of concentration versus log of distance. No visual evidence against linearity here. Again, adding a lowess smooth to this graph may enable a clearer assessment of possible of possible curvature.
If such displays lead to uncertain assessment, this may be an occasion to consult with a biostatistician who is familiar with interpretation of such displays. Researchers at the beginning of their use of such tools tend to 'see' too much in these sorts of displays. This can be thought of as a criticism of visual assessment. I do not agree. Some researchers feel the need to add pointless 'goodness-of-fit' tests

arguing that visual assessment is too 'subjective' and that , somehow, the GOF tests are then preferred. Nonsense.