REGRESSION, REALISTIC AND RELEVANT

JAN DE LEEUW

ABSTRACT.

1. REGRESSION

In regression we *smooth* an observed vector of outcomes y by replacing it by $\hat{y} = Py$, where P is a projector. Usually $P = X(X'X)^{+}X'$, with X a matrix of *predictors*. Most users of regression analysis are more interested in the *regression coefficients* $\hat{b} = (X'X)^{+}X'y$, because regression coefficients can be discussed in causal terms. We avoid that can of worms by concentrating on $\hat{y}$, and thinking of it as a result of smoothing.

There are other reasons to avoid regression coefficients. They tend to be numerically unstable and their value depends heavily on the way the predictors are *normalized* (or, more generally, *expressed*). The $\hat{y}$ are coordinate-free, in the sense that replacing X by XT gives the same $\hat{y}$, no matter what the (non-singular) T is.

Another important component of any regression analysis is to come up with information about the *stability* of the computed quantities. This can be done in a number of different ways. For a fairly general framework, we refer to **?**, Chapter 1.

2. STATISTICS: LINEAR MODEL

Statisticians study the stability of the outcomes of regression analysis by assuming that the vector of n observations y is a *realization* of an n-dimensional random vector $\underline{y}$.

Moreover we assume that the expectation $\mu \overset{\Delta}{=} \mathbf{E}(\underline{y})$ satisfies $\mu = P\mu$ and the dispersion $\Sigma \overset{\Delta}{=} \mathbf{E}\{(\underline{y} - \mu)(\underline{y} - \mu)'\}$ satisfies $\Sigma = \sigma^2 I$. Thus the elements of $\underline{y}$ are uncorrelated and they all have the same variance. And the means satisfy the model, i.e. $\hat{\mu} = \mu$. Very often we also assume that $\underline{y}$ has a multivariate normal distribution, which we write as $\underline{y} \sim \mathcal{N}(\mu, \Sigma)$, Observe we have *only one realization* of the random variable $\underline{y}$. Once us statisticians have these assumptions out of the way, we start doing regression calculations on the hypothetical random variables, not on the observations. Thus we compute $\underline{\hat{y}} = P\underline{y}$, and we can now study the distribution of $\underline{\hat{y}}$ under the model. The expectation is clearly $P\mu = \mu$ and the dispersion matrix is $\sigma^2 P$.

Statisticians also compute the residuals $\underline{\hat{r}} = \underline{y} - \underline{\hat{y}} = Q\underline{y}$, where $Q = I - P$. Clearly $\mathbf{E}(\underline{\hat{r}}) = 0$ and

$$\mathbf{E}(\underline{\hat{r}\hat{r}'}) = Q\mathbf{E}(\underline{yy'})Q = Q(\sigma^2 I + \mu\mu')Q = \sigma^2 Q,$$

and for the sum of squares of the residuals (or the *residual sum of squares*) we have

$$\mathbf{E}(\underline{\hat{r}}'\underline{\hat{r}}) = \sigma^2(n-p),$$

where $p = \mathbf{rank}(P) = \mathbf{tr}(P)$. Moreover

$$\mathbf{E}(\underline{\hat{r}\hat{r}'}) + \mathbf{E}\{(\underline{\hat{y}} - \mu)(\underline{\hat{y}} - \mu)'\} = \sigma^2 Q + \sigma^2 P = \sigma^2 I = \mathbf{E}\{\underline{y} - \mu)(\underline{y} - \mu)'\}$$

If $\underline{y}$ is normal, then so are $\underline{\hat{y}}$ and $\underline{\hat{r}}$. Moreover $\underline{\hat{r}}'\underline{\hat{r}}/\sigma^2$ and $(\underline{\hat{y}}-\mu)'(\underline{\hat{y}}-\mu)/\sigma^2$ are independent central chi-squares with $n-p$ and $p-1$ degrees of freedom, and these facts can be used to derive t and F distributions for various ratios.

This is all excruciatingly beautiful and the calculations seem to be easy to understand. But what does it all mean ? Where did the real world go ?

3. Statistics: Frequentists

What do we mean if we assume that our observations are realizations of random variables ? Frequentists interpret this as meaning that if we (hypothetically) replicate our experiment an infinite number of times, then the different independent realizations of the vector $\underline{y}$ will have a (normal) distribution with mean $\mu = P\mu$ and dispersion matrix $\Sigma = \sigma^2 I$.

Observe that *actual* replications are not directly relevant for this model. Such replications do not provide different realizations of the same random variable, they provide realizations of additional random variables with the same distribution as the previous one.

Three obvious remarks must be made here. In the first place, there is no way in which we can possibly verify this model, even approximately. We only have one realization of our multinormal random variable, and the hypothetical infinite *replication framework* is just that, hypothetical. It cannot be realized, so it cannot be tested. In short, it's metaphysics.

The second remark is that even if, by adopting some godlike features, we could observe the hypothetical replication framework, we would undoubtedly find that the model for this replication framework was radically and unrepairably wrong. If we do the thought experiment of replicating our study a large number of times, in our heads, then in almost all conceivable situations believing that the outcome will be what the linear model prescribes is just plain silly.

And then, in the third place, there are many experiments, especially in the social and behavioral sciences, that cannot be replicated at all, not even hypothetically in our heads. There is no conceivable framework of replication, because we study a situation which is essentially unique.

4. Statistics: Bayesians

5. Statistics: Minimal Model

	y	z	$\hat{y}$	$\hat{z}$
y	0			
z	2Σ	0		
$\hat{y}$	$\boxed{Q\Sigma Q + Q\mu\mu' Q}$	$\Sigma + P\Sigma P + Q\mu\mu' Q$	0	
$\hat{z}$	$\Sigma + P\Sigma P + Q\mu\mu' Q$	$Q\Sigma Q + Q\mu\mu' Q$	$2P\Sigma P$	0
μ	Σ	Σ	$P\Sigma P + Q\mu\mu' Q$	$P\Sigma P + Q\mu\mu' Q$

	y	z	$\hat{y}$	$\hat{z}$
y	0			
z	$2\sigma^2 I$	0		
$\hat{y}$	$\boxed{\sigma^2 Q + Q\mu\mu' Q}$	$\sigma^2(I + P) + Q\mu\mu' Q$	0	
$\hat{z}$	$\sigma^2(I + P) + Q\mu\mu' Q$	$\sigma^2 Q + Q\mu\mu' Q$	$2\sigma^2 P$	0
μ	$\sigma^2 I$	$\sigma^2 I$	$\sigma^2 P + Q\mu\mu' Q$	$\sigma^2 P + Q\mu\mu' Q$

	y	z	$\hat{y}$	$\hat{z}$
y	0			
z	$2\sigma^2 I$	0		
$\hat{y}$	$\boxed{\sigma^2 Q}$	$\sigma^2(I+P)$	0	
$\hat{z}$	$\sigma^2(I+P)$	$\sigma^2 Q$	$2\sigma^2 P$	0
μ	$\sigma^2 I$	$\sigma^2 I$	$\sigma^2 P$	$\sigma^2 P$

6. Stability

6.1. **Error Analysis.** Suppose we study the effect of uncertainty in the outcomes y_i. One way to do this is to codify uncertainly as the inequalities $y_i^- \leq y_i \leq y_i^+$. The question is how the $\hat{y}_i$ vary if the y_i vary in these *uncertainty intervals*. Now using $\hat{y} = Py$ we see that

$$\sum\{p_{ik}y_k^- \mid p_{ik} > 0\} + \sum\{p_{ik}y_k^+ \mid p_{ik} < 0\} \leq \hat{y}_i \leq \\ \sum\{p_{ik}y_k^+ \mid p_{ik} > 0\} + \sum\{p_{ik}y_k^- \mid p_{ik} < 0\}.$$

These are easily computed upper and lower bounds for the predicted values.

UCLA Department of Statistics
E-mail address: deleeuw@stat.ucla.edu