Ordinary Linear Regression - An historical comment and a way forward -

Rick Danielson, Will Perrie, Bash Toulany, Minghong Zhang, and Zhenxia Long
Fisheries and Oceans Canada, Bedford Institute of Oceanography, Dartmouth, Nova Scotia, Canada

Way Forward

A regression model is a "truth model" as much as it is an "error model", but our need to accommodate (for the first time) unshared truth *and* cross-correlated errors is not matched by an existing framework for doing so. Moreover, it is surprising that **only *four* extra samples surrounding each collocation seem to provide a tractable errors-in-variables solution.**

Nonlinear covariance

More flexibility may be needed in the physical interpretation of cross-correlation (or covariance) between two datasets. Although covariance is often taken to be a linear measure of association, it can be shown that an accommodation of correlated equation error yields an updated covariance between, say, in situ (*I*) and model (*N*) estimates:

$$Cov(I,N) = \beta\sigma_t^2 + \lambda_N\sigma_I^2.$$

Here, β is regression slope, σ_t^2 is shared true variance, σ_I^2 is error variance (say, in *I*) and λ_N is the fraction (between zero and one) of this error variance that is shared (hence, by *N*). **Regardless of the sign of β, we note that $\lambda_N\sigma_I^2$ is a *positive* error cross-covariance.** Although the traditional covariance of ordinary linear regression is defined by $\beta\sigma_t^2$ alone, a more sophisticated interpretation includes $\lambda_N\sigma_I^2$, which allows for a genuine correlation between two datasets that otherwise might be expected to have none.

Predictive Sampling

The new model employs a sampling strategy that *surrounds* the collocation of two datasets, say in space or in time (*T0*):

$$\begin{aligned}
\text{in situ }(T_0) \quad I &= t + \epsilon_I \\
\text{nowcast }(T_0) \quad N &= \alpha_N + \beta_N t + \lambda_N\epsilon_I + \epsilon_N \\
\text{forecast }(T_0 - 06h) \quad F &= \alpha_F + \beta_F t + \lambda_F(\lambda_N\epsilon_I + \epsilon_N) + \epsilon_F \\
\text{extended forecast }(T_0 - 12h) \quad E &= \alpha_E + \beta_E t + \lambda_E(\lambda_F(\lambda_N\epsilon_I + \epsilon_N) + \epsilon_F) + \epsilon_E \\
\text{revcast }(T_0 + 06h) \quad R &= \alpha_R + \beta_R t + \lambda_R(\lambda_N\epsilon_I + \epsilon_N) + \epsilon_R \\
\text{extended revcast }(T_0 + 12h) \quad S &= \alpha_S + \beta_S t + \lambda_S(\lambda_R(\lambda_N\epsilon_I + \epsilon_N) + \epsilon_R) + \epsilon_S.
\end{aligned}$$

For this model, called INFERS, the bracketing forecast and revcast samples constitute a timeseries that is ordered as *EFNRS*. (Note that by any definition of a forecast, the revcast is its functional inverse.) We employ *persistence* forecasts and revcasts, where samples are taken at equal time lags on opposite sides of a collocation. **Additive (α) and multiplicative (β) bias terms allow for a linear calibration of one dataset with respect to the other, and in the case of *FERS*, this calibration is also *predictive*.**

A fundamental justification for including shared error, following Mahalanobis (1947), Fuller (2006), and Janssen et al. (2007), is that individually, *I* and *N* are each expressed as linearly related to shared truth (*t*). Hence, although the equations for *I* and *N* are also an expression of the standard errors-in-variables regression model, they include an additional error term ($\lambda_N\epsilon_I$ is implicit in *I*, and thus is shared), which we identify as cross-correlated equation error.

Experimental Solution

A method-of-moments solution of INFERS has been proposed (Danielson et al., 2018) in which all equations for variance and for covariance that involve *I* or *N* are exact (i.e., equations for the covariance between *FERS* are solved as closely as possible. A simultaneous, weakly constrained minimization of differences between the LHS and RHS of six autocovariance equations (i.e., the covariances involving the forecast and revcast samples *FERS*) is sought. The proposed solution is tractable by identifying the two free parameters - shared true variance σ_t^2, and multiplicative calibration or regression slope, β_N - that are required to obtain a unique solution of the model.

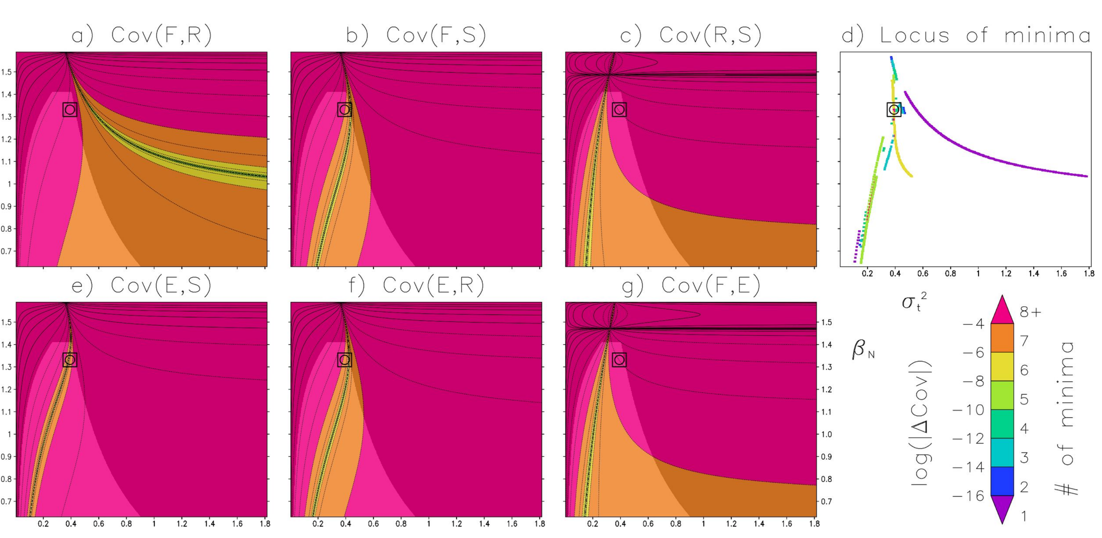

Figure 2: Demonstration of a weakly constrained, general solution of INFERS. This involves full parameter searches between zero and twice the variance of *I* for shared true variance (abscissa) and between OLS and RLS solutions for regression slope (ordinate). The open square denotes a prefered solution by a maximum (d) in the average position of autocovariance minima (a-c,e-g). The open circle denotes a final solution (i.e., closest to the open square, but constrained to the lighter shading).

Figure 2 is a typical example of localized paths (one for each equation) along which autocovariance differences are small. These pathways converge to a similar point at the top of each panel (the reverse linear regression solution). The unshaded region is where shared error fraction (λ_N) is between zero and one and all error variances are nonnegative. These constraints often exclude the reverse linear regression solution, although our selected solution (open circle) is not far away. This solution is chosen by identifying pathways in slices of fixed σ_t^2 or β_N. At the average location of all pathways in a slice, the number of contributing pathways is recorded. **All slices are summed to obtain a locus of minima (Fig. 1d), from which a *unique* solution is selected** (i.e., as a global maximum, possibly after smoothing if there are multiple local maxima).

Modern References

Carroll, R. J., and D. **Ruppert**, 1996. The use and misuse of orthogonal regression in linear errors-in-variables models. The American Statistician, 50 , 1–6.
Fuller, W. A., 2006. Errors in variables. Encyclopedia of Statistical Sciences, S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic and N. L. Johnson, Eds., doi:10.1002/0471667196.ess1036.pub2.
Janssen, P. A. E. M., Abdalla, S., Hersbach, H., and Bidlot, J.-R., 2007. Error estimation of buoy, satellite, and model wave height data. J. Atmos. Oceanic Technol., 24 , 1665-1677, doi:10.1175/JTECH2069.1.
Su, C.-H., Ryu, D., Crow, W. T., and Western, A. W., 2014. Beyond triple collocation: Applications to soil moisture monitoring. J. Geophys. Res. Atmos., 119, 6419-6439, doi:10.1002/2013JD021043.

Historical Comment

From Wikipedia (2019): "Six blind elephants were discussing what men were like. After arguing they decided to find one and determine what it was like by direct experience. The first blind elephant felt the man and declared, 'Men are flat.' After the other blind elephants felt the man, they agreed. Moral: We have to remember that **what we observe is not nature in itself, but nature exposed to our method of questioning**."

Figure 1:The parable of **six blind men** is over 2500 years old; the "inverse" parable of **six blind elephants** is 60 years old (Wikipedia 2019). Both parables refer to **partial truth** but how does this relate to linear regression?

A model for ordinary linear regression, and measurement models in general, are easier to formulate by assuming that a reference exists for absolute, genuine, or wholistic truth. However, this may come at an interpretive cost. **In ordinary linear regression, each datum is expressed as the sum of just *two* components: a deterministic term that is linearly related to the shared explanatory variable (called truth) and an independent, random term (called error).** Contrary to parables above, a single truth is wholly shared and separate errors are wholly unshared. What is the cost of neglecting notions of a) unshared or partial truth and b) dependent or cross-correlated error? More generally, what about shared and unshared components of both truth and error?

Regarding shared error, can we assume that our data are *linearly* related to truth? Janssen et al. (2007) note that "if there is actually a nonlinear relation between measurement and the truth but the linear calibration model ... would be taken instead, the error will have a random and a systematic component. Furthermore, if two types of measurements have a similar nonlinear relation with the truth,then in the context of the linear model ... there is now the possibility of correlated errors." By questioning whether two datasets are each linearly related to truth, we are not just motivating the introduction of a term called equation error (Carroll & Ruppert 1996, Fuller 2006). **By fundamentally questioning the linear assumption, we motivate a more elaborate measurement model in which each datum is expressed as the sum of *three* components: the more familiar shared truth and random error components, plus a third component that involves equation error**. Emphasis on shared error (nonlinear sharing), as well as shared truth (linear sharing), can be made explicit.

Might we be playing the blind elephant? Pearson (1902) was sensitive to error dependence but was challenged by unexpected correlations that were described as either spurious or genuine, with the identification of genuine correlation (taken as the result of physical or environmental similarities) being most challenging. Mahalanobis (1923) continued to experiment with cross-correlation based on estimated errors in tropospheric sounding measurements. Subsequently, Mahalanobis (1940, 1947) performed general experiments with measurements made using nonlinear reference scales. Following Pearson (1902), the hypothesis was entertained that subtle similarities among particular groups of measurements may have been present (and perhaps could be avoided). However, **Mahalanobis (1947) also acknowledged that this placed an emphasis on (as yet unidentified) physical or environmental similarities among groups of observers, rather than on the question of how measurements are first made and then later interpreted using a model.**

The growing emphasis on equation error seems to have begun after the 1940s. Mahalanobis (1947) highlights early experiments using steel rulers constructed with unequal measurement intervals (Mahalanobis, 1940). Nonlinear measurements using Kater pendulums at the National Bureau of Standards were proposed at the Allied Mathematics Colloquium Series lecture in November 1946, but it seems unlikely that Mahalanobis (1947) was successful in motivating them. Although Kruskal (1988) cites extensive research on repeated measurements into the 1980s, his description of the nonlinear reference scale suggests that such experiments are rarely considered. **A more elaborate measurement model might make it easier to connect theory and experiments**.

Historical References

Kruskal, W., 1988. Miracles and statistics: The casual assumption of independence. J. Amer. Statist. Assoc., 83 , 929-940.
Mahalanobis, P. C., 1923. On errors of observation and upper air relationships. Memoirs Indian Meteor. Dept., 24 , 11–19.
Mahalanobis, P. C., 1940. Errors of observation in physical measurements. Science and Culture, 5 , 443–445.
Mahalanobis, P. C., 1947. Summary of a lecture on the combination of data from tests conducted at different laboratories (reported by J. Tucker Jr.). Amer. Soc. Test. Materials Bulletin, 144 , 64–66.
Pearson, K., 1902. On the mathematical theory of errors of judgement, with special reference to the personal equation. Phil. Trans . Roy. Soc. London, 198 , 235–299.
Wikipedia, 2019. https://en.wikipedia.org/wiki/Blind_men_and_an_elephant, with quoted text from Heisenberg, Werner (1958). Physics and philosophy: the revolution in modern science. Harper. p. 58.

Ocean Applications

An ongoing challenge in most scientific domains is to characterize (dis)agreement between two methods of measurement. Predictive sampling may provide a new opportunity to resolve bias and performance when multiple samples are available near each collocation (cf. Bland and Altman 2007). The prospect of measurement model solutions that use these samples as "instruments" (Su et al., 2014, Danielson et al. 2018) can encourage a growing familiarity with shared truth and shared error as updated measures of linear and nonlinear agreement, respectively. Selected ocean applications are highlighted below.

Ecosystem models

Apart from ecosystem proxy variables that are regularly observed from space, in situ surveys are usually conducted a few times a year and historically target selected species. No attempt to compare observations and a growing list of numerical ecosystem models by predictive sampling (say, using five consecutive months) has yet been made. Although existing allometric theory is an attractive basis for many model components (Arhonditsis et al. 2019), an accommodation of *nonlinear* associations (covariance) would be timely.

Surface Visibility

Regional climate simulations depict a trend in relative humidity associated with decreased ice coverage in the Arctic in coming decades. Ship traffic is expected to increase but support services in case of accidents is a known problem. Associated with an increase in relative humidity is a decrease in visibility, but the **variance in fog and visibility that can be explained by relative humidity is poorly known**.

Danielson et al. (2019a,b) compare visibility observations from ships in the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) to proxies taken from the European Centre for Medium-Range Weather Forecasting Reanalysis (ERA) Interim dataset. **Presumably, shared error may include spurious correlation (Pearson 1902) as well as nonlinearity in the ICOADS-ERA visibility measures**. Shared truth alone is employed as the metric of agreement. ERA visibility proxies include three tuned parameterizations (GM, RUC, FSL), relative humidity (RHU), SST, specific humidity (SHU) and dew point temperature (DPT).

ICOADS	ERA Interim						
Unperturbed	GM	RUC	FSL	RHU	SST	SHU	DPT
Shared Truth	42.7	38.1	42.5	38.1	5.6	10.2	7.6
Shared Error	11.0	7.6	7.6	7.6	1.2	0.1	1.1
Unshared	46.3	54.2	49.9	54.2	93.2	89.7	91.3
Perturbed	(Arctic baseline)						
Shared Truth	38.1	32.9	34.7	32.9	4.4	10.7	10.1
Shared Error	11.5	13.1	16.9	13.1	2.8	0.2	1.0
Unshared	50.4	54.0	48.4	54.0	92.8	89.0	88.9
Perturbed	(with Neural Network)						
Shared Truth	41.4	41.5	41.7	41.5	2.4	6.1	6.3
Shared Error	7.3	7.2	7.3	7.1	2.2	1.5	3.5
Unshared	51.3	51.3	51.1	51.4	95.4	92.4	90.2

Table 1: Assessment of the linear relationship between ICOADS and ERA Interim visibility for 59966 Arctic collocations. *All values shown are a percentage of ICOADS variance*.

The ICOADS-ERA collocations are divided so that each ERA proxy can be trained to yield ICOADS visibility. **Training succeeds in showing that shared-error-nonlinearity can be reduced, as expected, but not to zero.** Shared error is thus interpreted as a combination of equation error and spurious correlation. The unshared category has a *complex* interpretation as measurement error as well as unshared components of equation error and truth(!)

Surface Current

A full solution of the INFERS model (Fig. 2) was only implemented in 2018. Hence, its application to surface current data can be compared to a partial solution given by Danielson et al. (2018). Unsurprisingly, their assumption of matching variance (and total error) does not carry into the full solution: GlobCurrent (nowcast) SNR is better, drifter (in situ) error is larger, and shared error is smaller. However, the full solution seems as robust as the earlier approximate solution.

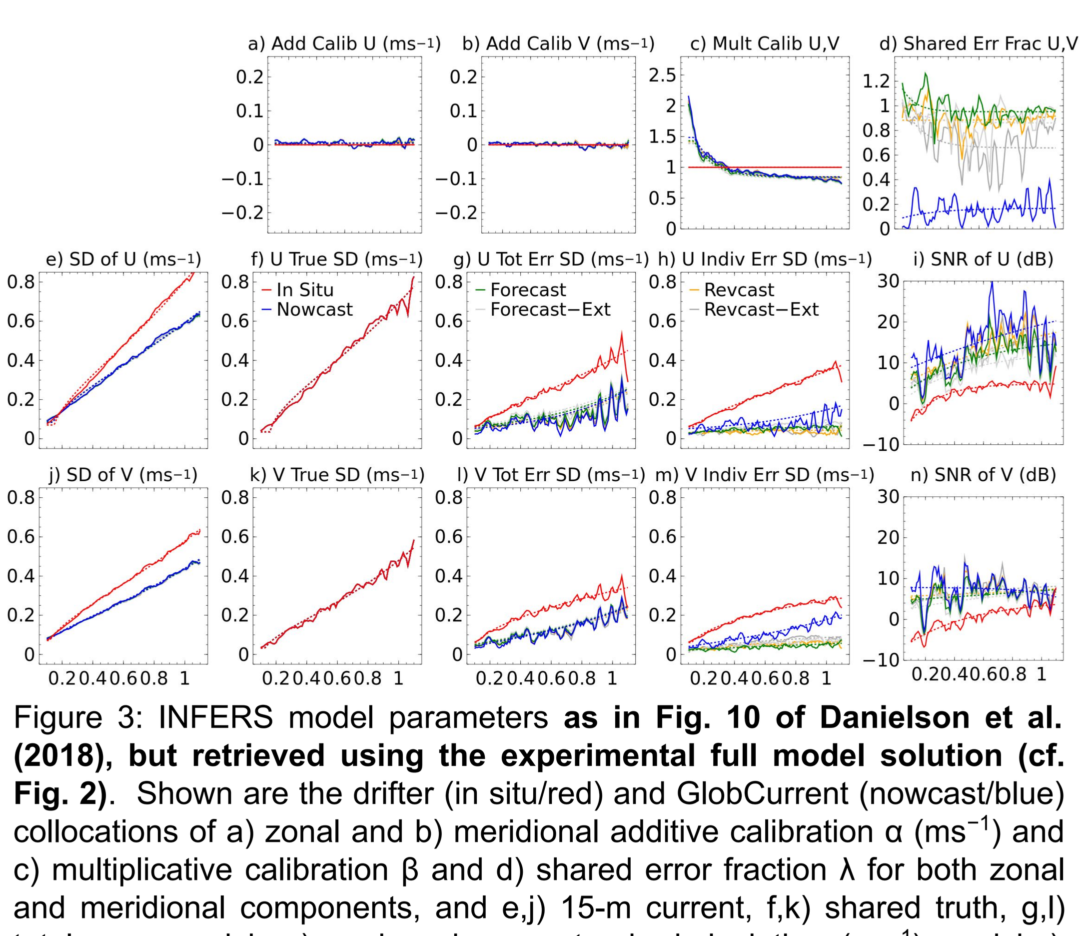

Figure 3: INFERS model parameters **as in Fig. 10 of Danielson et al. (2018), but retrieved using the experimental full model solution (cf. Fig. 2)**. Shown are the drifter (in situ/red) and GlobCurrent (nowcast/blue) collocations of a) zonal and b) meridional additive calibration α (ms^{-1}) and c) multiplicative calibration β and d) shared error fraction λ for both zonal and meridional components, and e,j) 15-m current, f,k) shared truth, g,l) total error, and h,m) unshared error standard deviation (ms^{-1}), and i,n) signal to noise ratio (dB) for the zonal and meridional components, respectively.

Applied References

Arhonditsis, G. B., Y. Shimoda, and N. E. Kelly, 2019. Allometric theory: Extrapolations from individuals to ecosystems, Encyclopedia of Ecology (Second Ed.), B. Fath, Ed., Elsevier, 242-255, doi:10.1016/B978-0-12-409548-9.10576-7.
Bland, J. M. and D. G. **Altman**, 2007. Agreement between methods of measurement with multiple observations per individual, J. Biopharm. Stats., 17:4, 571-582, doi:10.1080/10543400701329422.
Bentamy, A., Piollé, J.-F., Grouazel, A., Danielson, R. E., Gulev, S. K., Paul, F., Azelmat, H., Mathieu, P.-P., von Schuckmann, K., Sathyendranath, S., Evers-King, H., Esau, I., Johannessen, J. A., Clayson, C. A., Pinker, R. T., Grodsky, S. A., Bourassa, M., Smith, S. R., Haines, K., Valdivieso, M., Merchant, C. J., Chapron, B., Anderson, A., Hollmann, R., Josey, S. A., 2017. Review and assessment of latent and sensible heat flux accuracy over global oceans. Remote Sens. Environ. 201, 196-218, doi:10.1016/j.rse.2017.08.016.
Danielson, R. E., Johannessen, J. A., Quartly, G. D., Rio, M.-H., Collard, F., Chapron, B., and Donlon, C., 2018. Exploitation of error correlation in a large analysis validation: GlobCurrent case study. Remote Sens. Environ., 217 , 476–490, doi:10.1016/j.rse.2018.07.016.
Danielson, R. E., Zhang, M., and Perrie, W. A., 2019a. Possible impacts of climate change on fog in the Arctic and Subpolar North Atlantic, submitted and in review.
Danielson, R. E., Perrie, W. A., and Zhang, M., 2019b. On partial explanations of variance in marine visibility observations, submitted and in review.