Modelling the effect of different ingredients on the strength of high-performance concrete

a report compiled for the completion of DATA2002

High performance concrete production is more complex compared to traditional concrete (cement, water and fine/course aggregates only) due to additional materials. An example is adding heavier materials to create stronger HPC for improved safety, longevity and economic optimisation. As concrete is ubiquitous and difficult to understand the best combination of concrete elements, we can use modelilng to map the relationship between mixtures and concrete compressive strength. The report begins by challenging the current industry used model proposed by Abrams. In 1918, he argued that concrete strength can be predicted based on water to cement ratio alone. We then attempt to propose a more complex model.

concrete | high performance concrete | HPC | multiple linear regression

Table 1. Varaibles

Component	Measure	Mean	Description
Cement [a]	kg/m^3	232.2	Portland Cement
Water [a]	kg/m^3	186.4	Tap water
Coarse aggregate [a]	kg/m^3	943.5	Crushed Rock
Fine aggregate [a]	kg/m^3	819.9	River Sand
Fly ash [b]	kg/m^3	46.4	Powerplant waste
Blast furnace slag [b]	kg/m^3	79.2	Iron production waste
Superplasticizer [b]	kg/m^3	3.5	Chemical admixture
Age [a]	kg/m^3	45.7	Curing time

[a] traditional concrete; [b] HPC additive

Introduction. The report attempt to answer the following questions: 1. Is concrete compressive strength (CCS) best explained by Abrams model or is a more complex model that incorporates other predictor variables better? 2. Which variables significantly contribute to predicting the strength of high performance concrete, and how?

Data set. The dataset was collected by Yeh (1998) and includes 1000+ concrete samples from 17 different sources (figure 1) with components that are known to improve CCS. Yeh (1998) removed approximately 300 samples with aggregates $>$ 20mm and/or special curing conditions. However, this selection criteria is not explained and may confound results.

Analysis. A correlation matrix was used to visualise relationships between variables (figure 2). No variables violated the assumption of collinearity (correlation coefficient $>$ 0.8 or an obvious pattern in the graph), so all were retained (Actini, 2011). As many relationships between CCS and independent variables were non-linear, a log(independent_varable + 1) transformations was applied (figure 3). Figure 4 shows that only age was transformed as other relationships could not be transformed adequately.

To find an initial reasonable model, backward selection from full model with manual stepwise removal of variables was used (AIC did not remove variables). From figure 4, superplasticizer was removed due to non-significant p-value and course/find agreegated removed. All other variables were retained as removal significantly reduced the R^2 value. Finally, as fy ash and slag showed a non-relationship with concrete strength, a new model was finalised without these variables.

The model was checked using **10-fold cross validation** to compare efficacy of four different models generated. From figure 5, we can see that ____ model is better due to smaller error terms.

Results. The results of the four models we ended up with are presented in Table 2. We assess the validity of the models based on:

The determination coefficient (R-squared) value, where higher values are preferable The RSME and MAE errors, where smaller errors are preferable The residuals plot, which informs us whether the assumption of homoskedacity is met, and the qqplot, which tells us whether the assumption of normality is met. A comparison of predicted and observed compressive strength for the different models Based on these criteria, if we look initially at the simple linear models that test each of the predictor variables individually, we see that the determination coefficient is very small for all of them. Regardless of the other criteria, we can conclude that these models are not sufficient to explain concrete strength and we can discard them. //A look at individual models on water, ash, slag, age and cement gives an R-squared of 0.08, 0.01, 0.017, 0.3, 0.25. Age and Cement gives a better prediction overall.//

This leaves us with three models to compare: our simple Abrams model tests the Abrams hypothesis that concrete strength can be predicted based on water and cement alone (1); our complex model generated through manual stepwise selection (2); and a selective model that, in addition to removing the variables excluded in the previous model, removes fly ash and blast furnace slag due to their non-linearity (3).

Moving on to our simple model (1), which tests the Abrams hypothesis that concrete strength can be predicted based on water and cement alone, we see that it explains on 31% of the variability in concrete strength, which is not very informative (Table 2). In addition, both the RSME and MAE errors are quite a bit larger than for our other models, which suggests it is not a good fit (Figure [errors]). However, looking at the residuals plot for this model, the spread is slightly better than for our more complex models (2 and 3), as the residuals are truly random for this model (Figure [resids]).

Our more complex model explains 81% of the variability in concrete compressive strength, the highest of any of our models (Table 2). In addition, it has the lowest error values out of all our models (Figure [errors]). These two factors indicate that it is a good fit. However, the residuals plot shows a slight fanning out from left to right along the x-axis, meaning that the assumption of homoskedacity is not necessarily being met. This is problematic, as it means our model will be vulnerable to overfitting in some areas and underfitting in others (ref).

Not only is the homoskedacity assumption not met in our complex model, but there is some indication that the relationship between concrete compressive strength and blast furnace slag and fly ash is not linear, as there is a high amount of variability in these two variables, which means the linearity assumption may also not be met. Because of this, we tried creating a third model also removed fly ash and blast furnace slag in addition to the three variables removed in our complex model. We found that this model only explained 67% of concrete strength (Table 2). This might be an acceptable value if the model was robust

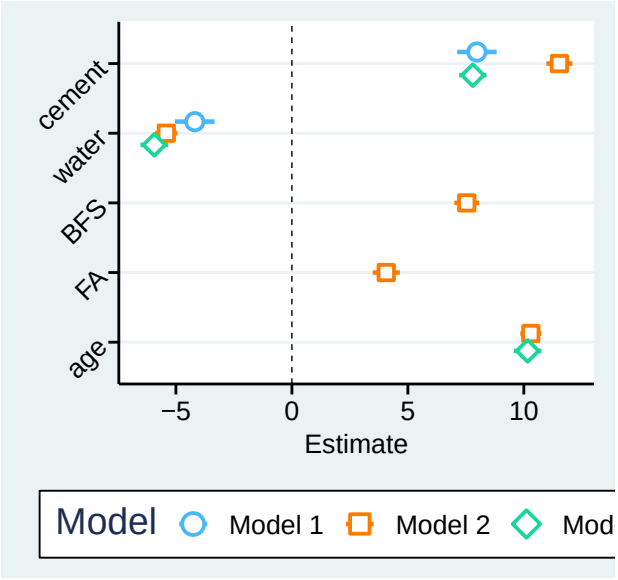


Fig. 1. Model estimate comparisons

in other ways; however, both the RSME and MAE errors for this model were larger than for model 2, and, although the spread of residuals appears slightly more random, it still suffers from the fan pattern, meaning homoskedacity is not satisfied in this model either.

Based on these parameters, we believe that the best model for predicting concrete compressive strength using multiple linear regression is:

$CS = 10.85 + 0.11(cement) + 0.09(BFS) + 0.06(FA) - 0.25(water) + 9.28(log(age+1))$

$CS = 10.85 + 0.11(cement) + 0.09(BFS) + 0.06(FA) - 0.25(water) + 9.28(log(age+1))$

$CS = 10.85 + 0.11(cement) + 0.09(\mathrm{BFS}) + 0.06(FA) - 0.25(water) + 9.28(log(age+1))$

Finally, we created plots which compared the actual values of concrete compressive strength to predicted values generated from the three different models (Figure XX) {fig2}. These visualisations suggest that our complex model generates predictions that best match the actual values. However, we can also see that as compressive strength increases, the model becomes less effective, with the points beginning to spread out. This is likely a consequence of the homoskedacity observed in the residual plot.

Discussion and Conclusion. Firstly, the Abramds model is not a sufficient model to predict CCS upon comparison with more complex models. This finding is suported by other research, which have similarly found that a more complex model is needed. Additionaly, the Abrams model is based in the 1918 production period were HPC are not as commonly found. Secondly, the five predictor variables (cement, age, water, slag and fly ash) were found to better predict CC based on the variable selection process.

Due to many roadblocks in this analysis, we believe multiple linear regerssion may not be an appropriate modelling approach for CCS. As (1) heteroskedacity was not solved (2) errors wih values _____ were still fairly high and (3) less effective model for higher CCS , it can be conclded that the model was the best between models but was not a good one. Additionally, there may be a risk of underfitting the model as some relationships between CCS and some IV (fly ash, slag, superplasticizer, fine and coarse agregates) were widely spread. Yeh (1998) came to the same conclusion with Artificial neural networks outperforming regression analysis. Despite regression analysis being used widely due to its simplicity (Atiili). Both Atili and Yeh argue that, due to the complexity of the concrete making process, it is not particularly effective, and we now have more powerful statistical tools such as artificial neural networks available that can produce more powerful models.

Some limitatinos of the analysis include: 1. unknown materials within superplasticisers 2. unknown fine and course agreggates 3. unknown mixing proportions and preparation techniques (Yeh 1998) 4. unreported class of fly ash 5. assumption that independent variables are independent when for example, increasing blast furnace slag requires reducing the amount of cement in a mixture (Atici 2011).

In future, this model can be improved by considering interaction between model variables and considering other analysis methods (e.g. ANN).[concluding sentence about the importance of concrete and importance of modelling].

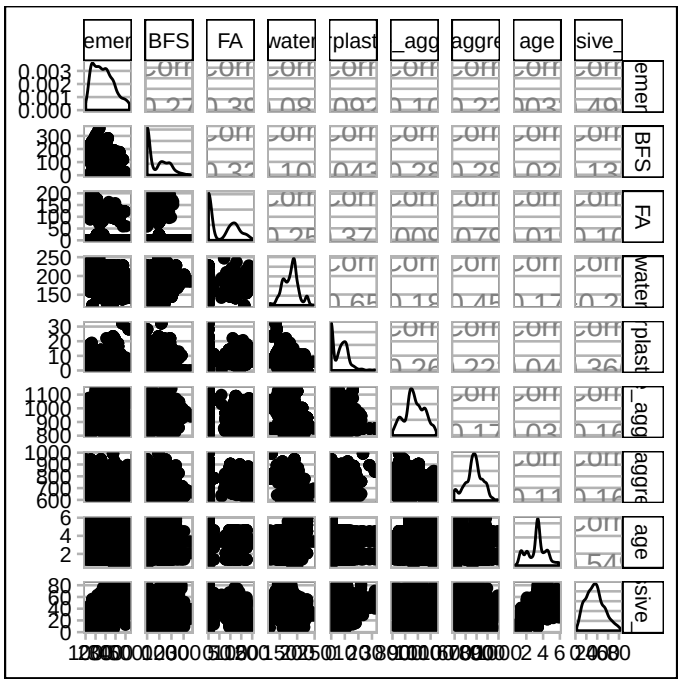
.

.

.

Appendix.

```
ggpairs(survey2) + theme_calc()
```



Single column equations.

$$\begin{aligned}(x+y)^3 &= (x+y)(x+y)^2 \\ &= (x+y)(x^2+2xy+y^2) \\ &= x^3+3x^2y+3xy^3+x^3.\end{aligned} \qquad [1]$$

Acknowledgments. This template package builds upon, and extends, the work of the excellent rticles package, and both packages rely on the PNAS LaTeX macros. Both these sources are gratefully acknowledged as this work would not have been possible without them. Our extensions are under the same respective licensing term (GPL-3 and LPPL (>= 1.3)).

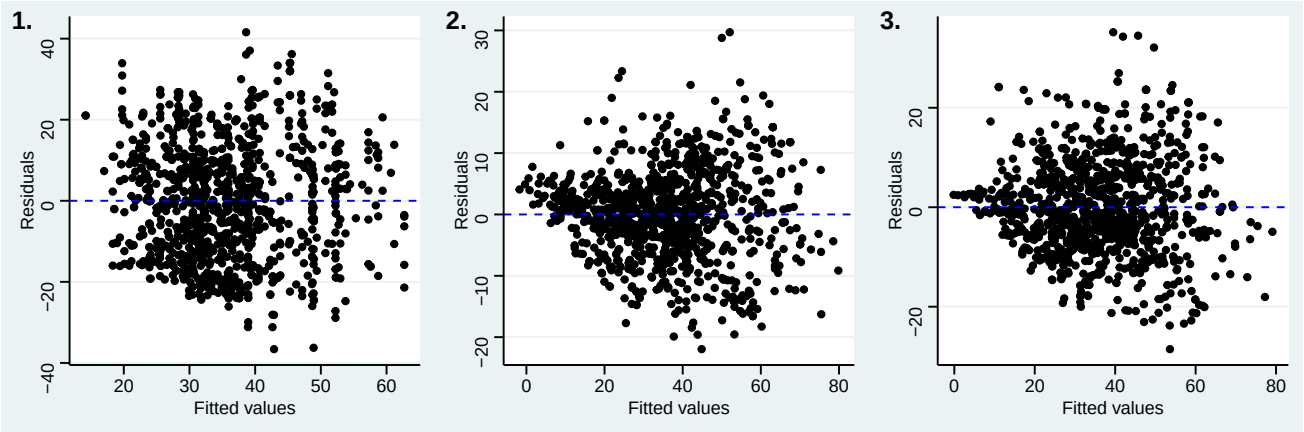


Fig. 2. Residuals for our three models: Simple Abrams model (1), Complex final model (2), and Selective model (3)

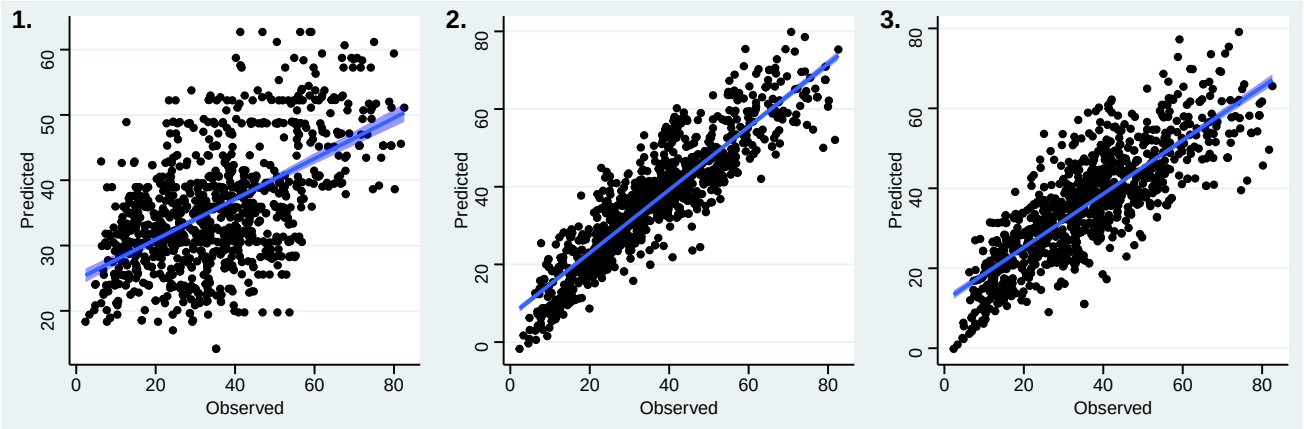


Fig. 3. Predicted vs observed values for our three models: Simple Abrams model (1), Complex final model (2), and Selective model (3)