December 17, 2018

To properly model fluid flow and calibrate the flow meters in our newest product, the StonePounder 2000, we require a highly accurate model for the viscosity of straight-chain hydrocarbons. Based on prior research, we feel confident that a model that uses the number of carbon atoms in the hydrocarbon will be sufficient as a predictor variable. A polynomial model was chosen to be the best because it is amenable to ordinary least squares regression (OLS), which is well known to produce the best modeling results. The experimental data we collected is as follows:

— So are many other functions

└ Ambiguous. If the OLS assumptions are met, it is BLUE (best linear unbiased estimator)

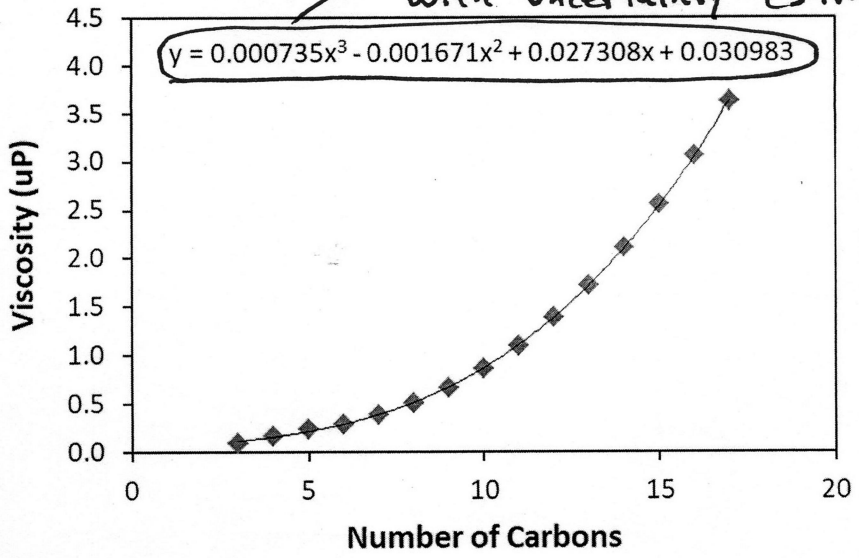
Name	# Carbons	Viscosity (uP)
propane	3	0.099
butane	4	0.168
pentane	5	0.245
hexane	6	0.296
heptane	7	0.39
octane	8	0.511
nonane	9	0.672
decane	10	0.863
undecane	11	1.1
dodecane	12	1.39
tridecane	13	1.72
tetradecane	14	2.11
pentadecane	15	2.56
hexadecane	16	3.06
heptadecane	17	3.61

Should include measurement uncertainty estimates.

(Measurements followed standard company procedure, as documented in my prior report.)

The data and the best-fit OLS model (3rd order polynomial) are presented in the following graph:

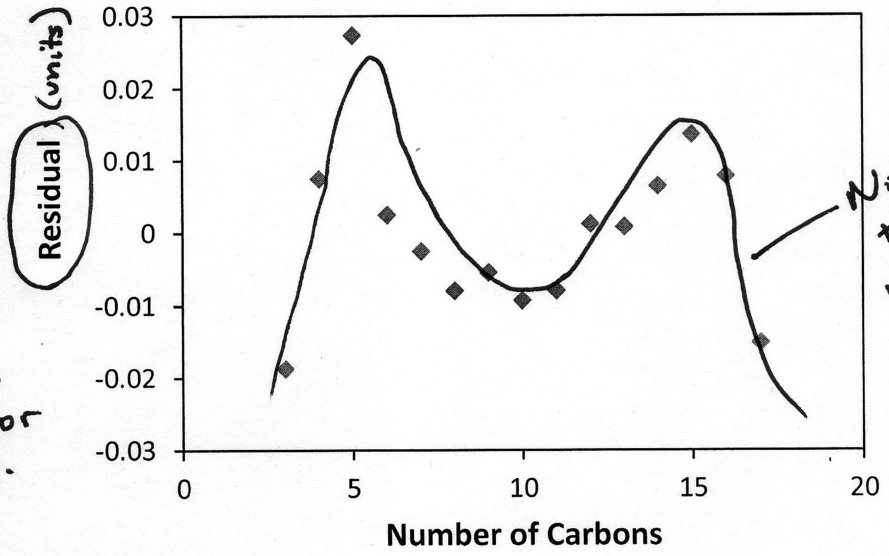
Model coefficients should be reported with uncertainty estimates.



x, x^2, x^3 will have huge multicollinearity.

Also, what is R^2 and standard deviation of residuals?

To test the appropriateness of the model, we first plot the residuals.



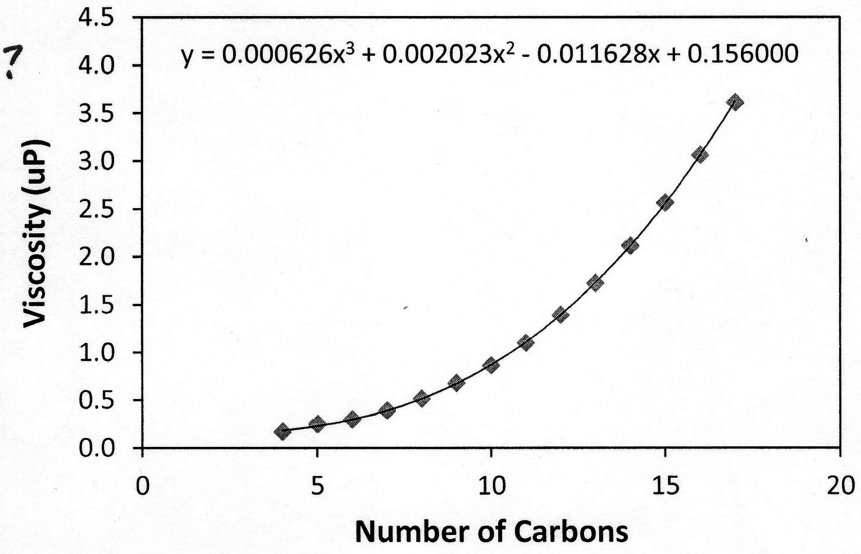
Better to plot esr (externally studentized residuals), though this is a minor point.

Note systematic trend indicating model error.

Attempts to use a lower order polynomial produced excessively high residuals, and higher order polynomials did not significantly improve the fit, so the third order polynomial was determined to be the best model to use. ← To compare models, use metrics like AIC, BIC, adjusted R^2, likelihood ratio, etc.

To determine the validity of the OLS model, some tests were performed on the residuals. A Shapiro-Wilks test on the residuals provided a p-value of 0.24, proving that the residuals were normally distributed. No statistical outliers were found, confirming the quality of the data used to calibrate the model.

To assess influence, the leverage and Cooks Distance were calculated for each residual. An influential data point was discovered, corresponding to propane, with high leverage and a large residual. As a result, that data point was removed from the data set and a new model was developed:



The large changes in the model coefficients with the propane data point removed shows that it is highly influential.

Based on these modeling results, I am confident that our new model will meet the requirements for its use in the StonePounder 2000.