MULTIPLE REGRESSIONS II

Today we explore data for 292 streamgages in the 48 contiguous US during the wateryear 2016 (from the USGS; https://is.gd/Ie9x0n). The raw data have already been organized to obtain one row per streamgage station. The data file (streams.csv) includes the following columns:

- STAID = streamgage station ID number
- DRAIN_SQKM = drainage area (km^2)
- PPTAVG_BASIN = average annual basin precipitation (cm)
- T_AVG_BASIN = average temperature in the drainage basin (ºC)
- RH_BASIN = average relative humidity in the basin (%)
- RRMEDIAN = median relief ratio (slope)
- **meanflow** = annual mean discharge (ft^3 / sec)

Your Mission: Find a multiple regression that

a) most plausibly (via AIC) predicts **meanflow** using listed predictors (& is more predictive than a null model). Your predictors need to make sense when you explain them.
b) represents hypotheses you make (e.g., meanflow is all about the basin / moisture / both, etc.)
c) best meets assumptions of normality and homogeneity of variance for residuals
d) avoids high collinearity among predictors (see `VIF` in the performance output)
e) provides a decent linear and tight scatter in `plot(meanflow,predict(your_best_model))`

You may need to transform meanflow.

How close to zero is your Intercept? Should it be ~ zero?

Which variable is most important to mean flow in your "best" model? It is hard to tell with such different units, so re-run your model using **`scale(X)`** where X is each predictor in your model. This uses a z-score to express every predictor as standard deviations, so you can compare them fairly. You still need the original if you want to talk about effect size in original units (e.g. mean flow increase one unit for every X km^2 of drainage area).