The correlation between CPU components and CPU performance

Executive summary/abstract
Address the connection between CPU hardware statistics and CPU performance. This report analyzes the computer hardware dataset in order to find a correlation between CPU components (covariates) with CPU performance (target) with a hierarchical linear regression model. The model convergence is archived and the result confirms that there is a strong correlation between CPU components and CPU performance.

Introduction
It is natural to think that there is a correlation between CPU components and CPU performance and different companies produce CPUs with different performance statistics. The research question in this report is to

- Confirm and approximate the amount of correlation.
- Validate that CPU performance varies between vendors.

This research question can be modeled with a hierarchical linear regression model with covariates are the CPU components statistics and the target is the relative CPU performance.

Data
The data used in this report is from:
https://archive-beta.ics.uci.edu/ml/datasets/computer+hardware

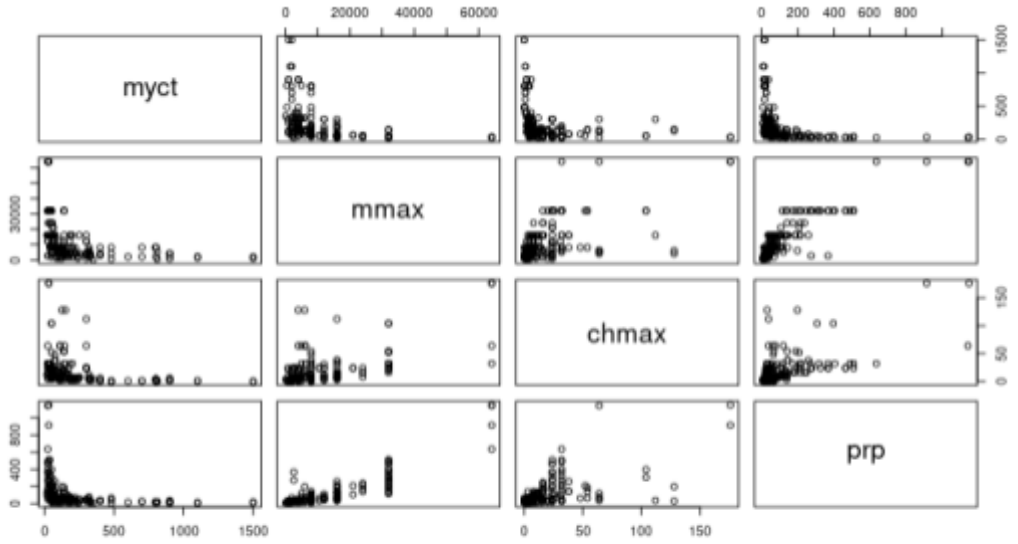
Each row contains the statistics of a CPU's components. There are a total of 209 CPU model instances, each instance have 10 attributes. We are only interested in the following 6 attributes:

- **vendor**: name of the company produce the CPU
- **myct**: machine cycle time in nanoseconds (integer)
- **mmax**: maximum main memory in kilobytes (integer)
- **cache**: cache memory in kilobytes (integer)
- **chmax**: maximum channels in units (integer)
- **prp:** published relative performance (integer) - Target

Below is some statistics and plots of the dataset

```
   vendor            myct            mmax           chmax           cache            prp
Length:209        Min.   :  17.0   Min.   :   64   Min.   :  0.00   Min.   :  0.00   Min.   :   6.0
Class :character  1st Qu.:  50.0   1st Qu.: 4000   1st Qu.:  5.00   1st Qu.:  0.00   1st Qu.:  27.0
Mode  :character  Median : 110.0   Median : 8000   Median :  8.00   Median :  8.00   Median :  50.0
                  Mean   : 203.8   Mean   :11796   Mean   : 18.27   Mean   : 25.21   Mean   : 105.6
                  3rd Qu.: 225.0   3rd Qu.:16000   3rd Qu.: 24.00   3rd Qu.: 32.00   3rd Qu.: 113.0
                  Max.   :1500.0   Max.   :64000   Max.   :176.00   Max.   :256.00   Max.   :1150.0
```

Dataset statistic
The predictive attributes is heavily right-skewed



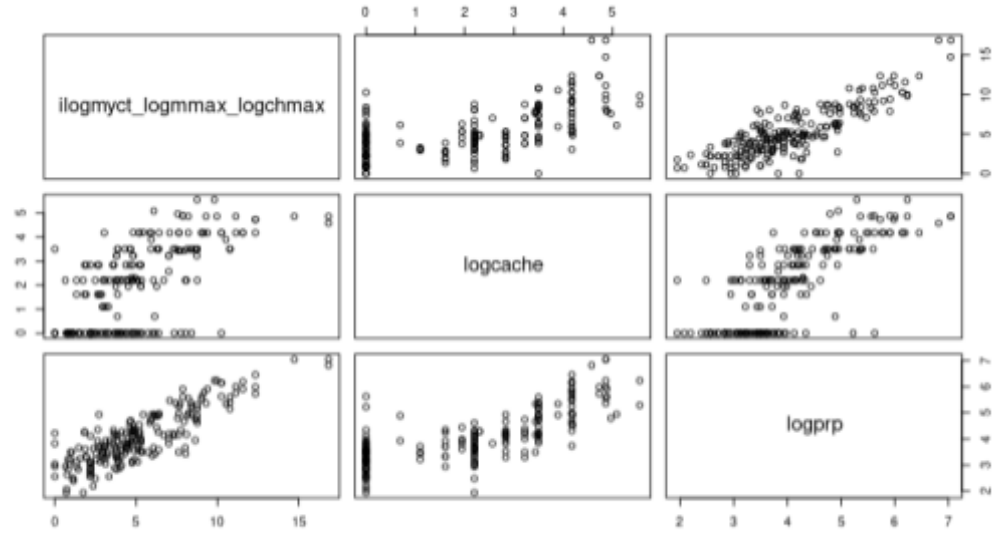
Pairs plot of attributes

The above pair plot show some level of correlation between **mmax, chmax** and **prp**, however the correlation is not clear as the distribution of each attribute is right-skwed and their range is different. In order to mitigate this effiect, the log transform is used to normalize the dataset.

Model

In order to model the correlation between CPU components and CPU performance, we will used the following covariates:

- **ilogmyct_logmmax_logchmax:** This covariate model the interaction between the 3 attribute: the inverse log cycle time, the log maximum main memory and the log maximum channels in each CPU. Those attribute together influence the CPU performance and we expected this covariate will have positive correlation with the CPU performance.
- **logcache**: the log cache memory.



Pairs plot of covariates

In order to model the variation of CPU performance between vendors, we will use a hierarchical random intercept model for the vendor group. The hierarchical linear regression will have the following form:

$$y_i \mid a_{vi},\ \theta,\ \phi,\ \sigma \sim N(\mu_i,\ \sigma),\ i = 1,\ldots,209$$
$$\mu_i = a_{vi} + \beta_0\,\theta + \beta_1\phi$$
$$a_v \sim N(0,\ 100),\ v = 1,\ldots,30$$
$$\beta_0,\ \beta_1 \sim N(0.5,\ 100)$$
$$\sigma \sim IG(0.5,\ 0.5)$$

y_i: $logprp$

a_v: $intercept\ for\ vendor\ v$

θ: $ilogmyct\_logmmax\_logchmax$

ϕ: $logcache$

The posterior distribution for each coefficient will be approximated by Gibbs Sampling algorithm.

Results

The trace plot of each coefficient and the Gelman and Rubin's convergence diagnostic confirm that the model have converged.

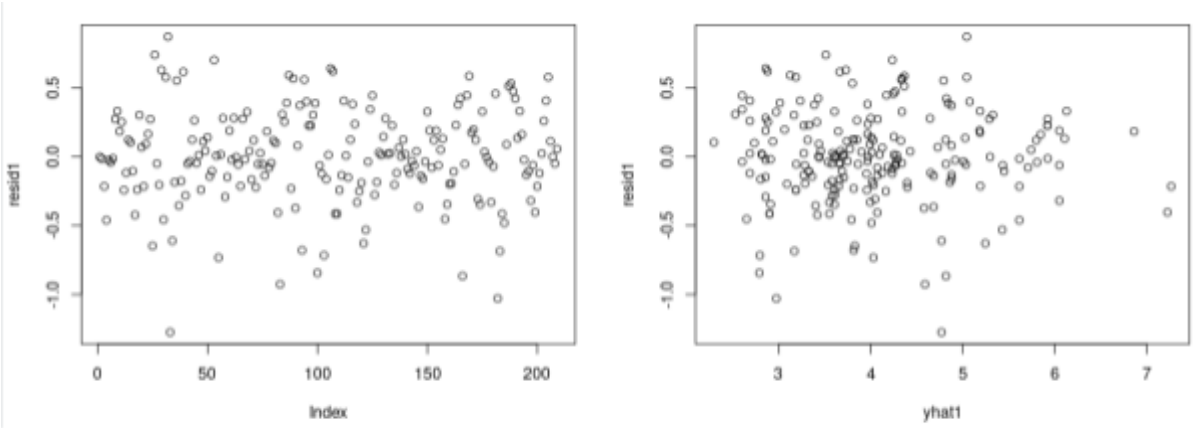
	a[1]	a[2]	a[3]	a[4]	a[5]	a[6]	a[7]	a[8]	a[9]	a[10]	a[11]
Lag 0	1.000000000	1.000000000	1.000000000	1.000000000	1.000000000	1.000000000	1.000000000	1.000000000	1.00000000	1.000000000	1.000000000
Lag 1	0.086428811	0.486782837	0.014749818	0.084968505	0.030308685	0.248555575	0.030810125	0.084098100	0.27520055	0.049179671	0.059531992
Lag 5	0.031862426	0.259495440	0.008828337	0.040527546	0.026827090	0.156449053	0.015655032	0.035566150	0.12358090	0.020447417	0.022174741
Lag 10	0.009234166	0.102579534	0.003710215	0.002115237	0.007646413	0.080033114	-0.004416726	0.015108141	0.03856498	0.003813988	0.018409820
Lag 50	0.005585100	0.005595724	-0.004328071	0.001868857	-0.001591983	0.007028294	0.002925636	0.001250879	0.01385758	0.002137998	0.001841279

	a[12]	a[13]	a[14]	a[15]	a[16]	a[17]	a[18]	a[19]	a[20]	a[21]	a[22]
Lag 0	1.000000000	1.000000000	1.000000000	1.00000000	1.00000000	1.00000000	1.00000000	1.0000000000	1.000000000	1.0000000000	1.000000000
Lag 1	0.031817103	0.029842491	0.156774654	0.12316640	0.22798908	0.18519564	0.26079292	0.1135585852	0.146794210	0.0849238823	0.454237539
Lag 5	0.020648595	0.005474223	0.071898410	0.07006391	0.12386773	0.10856390	0.13624723	0.0560523853	0.072556498	0.0503919992	0.198496002
Lag 10	0.012424134	0.002036750	0.024069741	0.03565849	0.05934273	0.04987676	0.06313661	0.0239105049	0.026686157	0.0358619544	0.052891534
Lag 50	-0.004842069	0.003696789	-0.006984723	0.01017904	0.01115976	-0.00477345	0.00696142	-0.0009398239	0.009119917	-0.0003727166	0.007037421

	a[23]	a[24]	a[25]	a[26]	a[27]	a[28]	a[29]	a[30]	b[1]	b[2]	sig
Lag 0	1.000000000	1.000000000	1.0000000000	1.000000000	1.00000000	1.00000000	1.000000000	1.0000000000	1.00000000	1.00000000	1.000000000
Lag 1	0.263465236	0.010855737	0.0397886589	0.087868044	0.29995863	0.40355356	0.028657727	0.0235948093	0.90881622	0.86837987	0.158432916
Lag 5	0.149289922	0.015142570	0.0197963025	0.033827753	0.12736390	0.21951270	0.019050852	0.0156603521	0.59671443	0.47950227	0.013994071
Lag 10	0.077494396	0.006783471	0.0115921205	0.012589084	0.04648927	0.10164694	0.013929621	0.0001457978	0.32000186	0.21311402	0.003310556
Lag 50	0.007965463	-0.002229608	0.0008582195	-0.004012492	0.01309455	0.01054926	0.006532869	-0.0035077903	0.01560644	0.01051335	-0.003126037

Autocorrelation in Markov chain

We can examine the fit of the model with the data by looking at the residual error plot:



Residual plots

The residual plots demonstrate a random pattern of noise between the model prediction and the grouthtruth, indicating that the defined model fits well with the data and has good predictive performance.

The posterior mean of each coefficient is given below

We can see that the intercept coefficient of each vendor is different, confirming that there is a variation in CPU performance between vendors. The posterior mean for the coefficient of the **ilogmyct_logmmax_logchmax** and **logcache** covariates is positive with some order of magnitude of the standard deviation, indicating that there is a positive correlation between the CPU components and CPU performance.

	Mean	SD	Naive SE	Time-series SE
a[1]	2.1812	0.67106	0.0038744	0.0051126
a[2]	2.2556	0.29860	0.0017240	0.0046936
a[3]	3.0246	0.45343	0.0026179	0.0026410
a[4]	2.7741	0.47291	0.0027303	0.0037215
a[5]	1.9913	0.45974	0.0026543	0.0034804
a[6]	2.0131	0.25909	0.0014959	0.0033380
a[7]	2.7984	0.26407	0.0015246	0.0017331
a[8]	2.2575	0.29958	0.0017296	0.0022944
a[9]	2.2575	0.25073	0.0014476	0.0028704
a[10]	2.7022	0.26871	0.0015514	0.0019420
a[11]	3.1111	0.24888	0.0014369	0.0018094
a[12]	2.2826	0.29244	0.0016884	0.0020553
a[13]	2.4577	0.65558	0.0037850	0.0042664
a[14]	2.4154	0.40525	0.0023397	0.0037226
a[15]	2.3857	0.25877	0.0014940	0.0023802
a[16]	2.1453	0.20217	0.0011672	0.0023275
a[17]	1.9327	0.26545	0.0015326	0.0028695
a[18]	2.6036	0.13105	0.0007566	0.0015382
a[19]	2.5144	0.27771	0.0016034	0.0024200
a[20]	1.8062	0.28565	0.0016492	0.0026437
a[21]	1.2955	0.67269	0.0038838	0.0061170
a[22]	2.7095	0.19851	0.0011461	0.0027206
a[23]	2.0248	0.20610	0.0011899	0.0026288
a[24]	2.6454	0.37532	0.0021669	0.0022898
a[25]	2.7766	0.38121	0.0022009	0.0026676
a[26]	2.5070	0.29612	0.0017096	0.0022773
a[27]	2.1981	0.22124	0.0012773	0.0025761
a[28]	2.0986	0.22909	0.0013227	0.0034204
a[29]	2.6330	0.65090	0.0037580	0.0044471
a[30]	3.7744	0.45809	0.0026448	0.0030265
b[1]	0.2658	0.02434	0.0001405	0.0005945
b[2]	0.1421	0.04529	0.0002615	0.0009671
sig	0.6399	0.03384	0.0001954	0.0002308