Separate the tissue term and life stage, then combine.

Selecting different subsets of all the genes, try different ones and validate the rest (cross-validation)

Look at both tissue and life stages, first separately and then together.

Look at the biological influence of these weights in different tissues, do they correlate to actually biological meanings.

Looking at the early stage and late stage, since late stage can consume food the energy efficiency should be less important thus the selection is weaker.

For each weight, there is a correlation, plot that.

Plug our model in the WLS model.

NMF, or traditional heatmap to seek if we are grouping the tissues correctly

Scale a list so the values in the list have an average of 1.

Look at the residuals (heteroskedastic)

Log likelihood at MLE (where chi-square test comes from)

Try it with intercept oF 1, the true fitted intercept should be 0 or close to 0.

These weights coming from our regression analysis may correlate with length, cell, ATP.

If we scale these using housekeeping genes, we can validate if our scaling is robust

Weighted least squares has several advantages over other methods, including:

It’s well suited to extracting maximum information from small data sets.

It is the only method that can be used for data points of varying quality.

Disadvantages include:

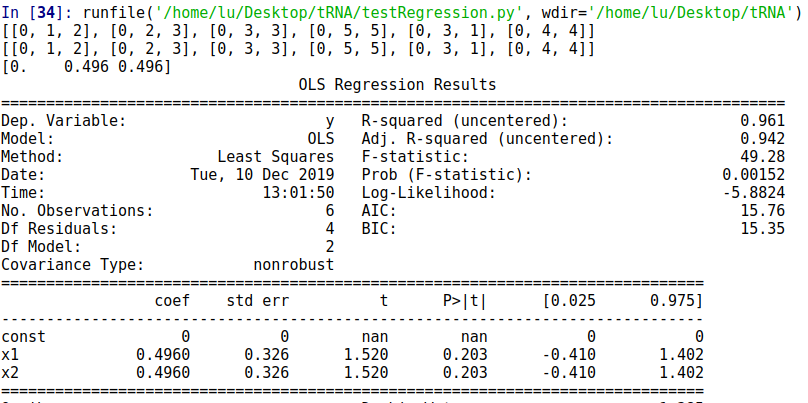
It requires that you know exactly what the weights are. Estimating weights can have unpredictable results, especially when dealing with small samples. Therefore, the technique should only be used when your weight estimates are fairly precise. In practice, precision of weight estimates usually isn’t possible.

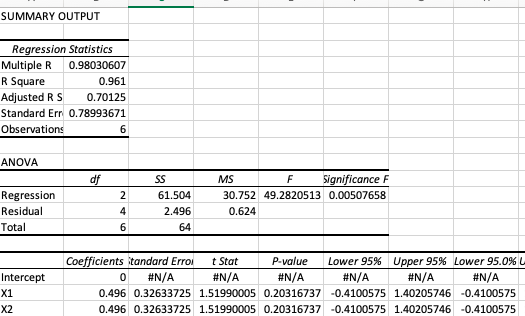
Sensitivity to [outliers](https://www.statisticshowto.datasciencecentral.com/find-outliers/) is a problem. A rogue outlier given an inappropriate weight could dramatically skew your results.

Each Correlation Measurement takes Roughly 0.007 sec, if we are to get them done around a minuted, we need to control them to be under 10000

Gamma distribution

Confirming the regression was done correctly:





|  |  |  |  |
| --- | --- | --- | --- |
| Species |  | const | Regression Coef |
| male, organism, embryo Ce |  | x25 | -0.694 |
| hermaphrodite, organism, newly molted young adult hermaphrodite Ce |  | x19 | -0.5839 |
| hermaphrodite, organism, L3 larva Ce |  | x10 | -0.221 |
| hermaphrodite, organism, post dauer stage Ce |  | x20 | -0.1589 |
| hermaphrodite, organism, late cleavage stage embryo Ce |  | x18 | -0.1176 |
| hermaphrodite, neuron, L1 larva Ce |  | x4 | -0.0947 |
| hermaphrodite, organism, proliferating embryo Ce |  | x21 | -0.0384 |
| hermaphrodite, organism, L1 larva Ce |  | x7 | -0.0322 |
| hermaphrodite, pharyngeal muscle cell, fully-elongated embryo Ce |  | x22 | -0.019 |
| hermaphrodite, organism, 3-fold embryo Ce |  | x5 | -0.0156 |
| hermaphrodite, gonad, adult Ce |  | x2 | -0.0109 |
| hermaphrodite, NSM, L1 larva Ce |  | x1 | -0.0082 |
| male, organism, L4 larva Ce |  | x24 | 0.0177 |
| hermaphrodite, organism, fully-elongated embryo Ce |  | x16 | 0.0291 |
| hermaphrodite, somatic cell, embryo Ce |  | x23 | 0.031 |
| hermaphrodite, motor neuron, L2 larva Ce |  | x3 | 0.0664 |
| hermaphrodite, organism, dauer larva Ce |  | x13 | 0.0688 |
| hermaphrodite, organism, enclosing embryo Ce |  | x15 | 0.1478 |
| hermaphrodite, organism, L2d-dauer molt Ce |  | x9 | 0.1932 |
| hermaphrodite, organism, adult Ce |  | x12 | 0.203 |
| hermaphrodite, organism, 4-cell embryo Ce |  | x6 | 0.2262 |
| hermaphrodite, organism, elongating embryo Ce |  | x14 | 0.2348 |
| hermaphrodite, organism, gastrulating embryo Ce |  | x17 | 0.274 |
| hermaphrodite, organism, L4 larva Ce |  | x11 | 0.342 |
| hermaphrodite, organism, L2 larva Ce |  | x8 | 0.6588 |

0.16354809452180194

0.23581812889159412

0.282132886991515

0.18009761677549993

0.25524644664536655

0.2864684271235175

0.30192990276284687

0.2904688903797501

0.2814600180522045

0.2116676828849915

0.1774797733534051

0.23469321254859915

0.22189441069085944

0.270745539297386

0.22166141257914582

==

0.3872889954730867

0.2500812326402075

0.3754411086625951

0.25824213548847497

0.27957414827578586

0.35720735409660925

0.3495400598765668

0.33857918870602

0.38973357157632704

0.32433117294640634

0.23768676468903882

0.21250234562148299

0.34396206848881683

0.2548944441327616

0.27004059635841243

0.32107070422702394

==

0.4963031368128077

WLS Regression Results

==============================================================================

Dep. Variable: y R-squared: 0.226

Model: WLS Adj. R-squared: 0.226

Method: Least Squares F-statistic: 414.1

Date: Thu, 02 Jan 2020 Prob (F-statistic): 0.00

Time: 12:57:01 Log-Likelihood: 17979.

No. Observations: 21252 AIC: -3.593e+04

Df Residuals: 21236 BIC: -3.580e+04

Df Model: 15

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

const 0.1665 0.001 232.265 0.000 0.165 0.168

x1 0.3840 0.057 6.691 0.000 0.272 0.497

x2 0.4201 0.044 9.502 0.000 0.333 0.507

x3 0.8327 0.046 18.079 0.000 0.742 0.923

x4 0.2011 0.037 5.486 0.000 0.129 0.273

x5 0.8493 0.062 13.798 0.000 0.729 0.970

x6 0.6990 0.090 7.788 0.000 0.523 0.875

x7 1.2095 0.122 9.953 0.000 0.971 1.448

x8 0.5112 0.049 10.375 0.000 0.415 0.608

x9 0.1290 0.114 1.129 0.259 -0.095 0.353

x10 -4.0737 0.217 -18.803 0.000 -4.498 -3.649

x11 1.3896 0.168 8.293 0.000 1.061 1.718

x12 0.5280 0.058 9.071 0.000 0.414 0.642

x13 0.5831 0.068 8.599 0.000 0.450 0.716

x14 -0.5086 0.091 -5.564 0.000 -0.688 -0.329

x15 -0.7940 0.128 -6.214 0.000 -1.044 -0.544

==============================================================================

Omnibus: 2312.190 Durbin-Watson: 1.472

Prob(Omnibus): 0.000 Jarque-Bera (JB): 8641.078

Skew: 0.515 Prob(JB): 0.00

Kurtosis: 5.949 Cond. No. 382.

==============================================================================

OLS Regression Results

==============================================================================

Dep. Variable: y R-squared: 0.247

Model: OLS Adj. R-squared: 0.246

Method: Least Squares F-statistic: 464.1

Date: Thu, 02 Jan 2020 Prob (F-statistic): 0.00

Time: 12:57:49 Log-Likelihood: 17331.

No. Observations: 21252 AIC: -3.463e+04

Df Residuals: 21236 BIC: -3.450e+04

Df Model: 15

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

const 0.1854 0.001 237.308 0.000 0.184 0.187

x1 0.3666 0.054 6.743 0.000 0.260 0.473

x2 0.5363 0.046 11.667 0.000 0.446 0.626

x3 0.8140 0.044 18.307 0.000 0.727 0.901

x4 0.1563 0.034 4.616 0.000 0.090 0.223

x5 0.8541 0.061 13.938 0.000 0.734 0.974

x6 0.7343 0.082 9.002 0.000 0.574 0.894

x7 1.2669 0.117 10.797 0.000 1.037 1.497

x8 0.4247 0.048 8.849 0.000 0.331 0.519

x9 0.1199 0.118 1.018 0.309 -0.111 0.351

x10 -4.2223 0.208 -20.327 0.000 -4.629 -3.815

x11 1.4160 0.162 8.755 0.000 1.099 1.733

x12 0.6314 0.061 10.285 0.000 0.511 0.752

x13 0.5219 0.065 8.059 0.000 0.395 0.649

x14 -0.4149 0.100 -4.146 0.000 -0.611 -0.219

x15 -0.9461 0.126 -7.507 0.000 -1.193 -0.699

==============================================================================

Omnibus: 3933.753 Durbin-Watson: 1.470

Prob(Omnibus): 0.000 Jarque-Bera (JB): 12394.642

Skew: 0.950 Prob(JB): 0.00

Kurtosis: 6.223 Cond. No. 344.

==============================================================================

Divide data into two, and predict one with another.

Rescale so the average is 1, like phi. Problem with that

PCA using log

Plot the fitted regression line and compare them with average and naive regression line

Ask alex if the std err or is std

Send them emails over the raw data.

If we get rid of the dauer specific genes, make a new set of genes.

Rerun ROC using this new gene, locking the mutation but different selection. So we will have the same delta M but different eta.

Key idea: do we see if dauer specific genes have a different codon usage bias.

If we can separate the genes are in dauer stages, we can do further clustering with probability functions to determine which genes go to which cluster.

Remove the genes and the regression that involve dauer stages. And do dauer vs non-dauer, we can do them for different stage, non-dauer stages.

It’s possible that other stages are responsible for this too.

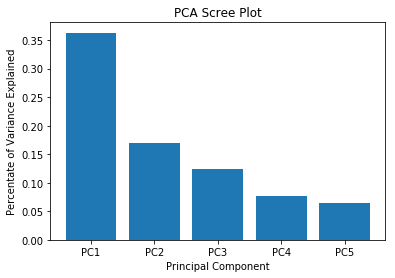
We could do dauer stage + housekeeping genes VS embryo stage genes, and expand those to all lifes stages (life stage genes + housekeeping genes), and we will have 17 of these (17 lifestages)

Scale a list so the average is 1, the problem

PCA log and non-Log:

[0.36293881 0.16992975 0.12369924 0.07644049 0.06522573]

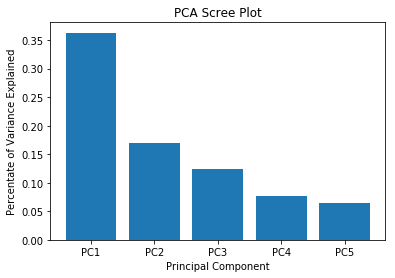
0.7982340104887727



[0.33297133 0.10937542 0.08702409 0.06464689 0.0499726 ]

0.6439903298719641

This seems to suggest that non-log explains more variation



Dauer-enriched genes are downregulated upon dauer recovery, showing either early or delayed kinetics. The set of dauer-enriched genes probably define those that confer dauer-specific properties, such as stress resistance and longevity. (wang 2003)

Highly expressed but most in dauer.

Take the genome, duplicate all the information.

Randomly split the genes in half, take ROC-SEMPPR and the gene expression values for each stage, each stage train ROC, set s-epsilon to 0.1. Take codon specific parameters and look at the other half, just plot phi.Take the previous fit and run ROC to estimate phi. Do that for every stage. Look at the delta eta and delta M between stages.

Choose the key stages out of 17. (Maybe 5)

If you look at the whole trace, it would seem that it flattened.

