BIOL-GA 2031. Statistics and Machine Learning in Genomics Homework 01

Prof. Manpreet S Katari
TA. Jaime Cascante Vega
Email. jc12343@nyu.edu

In the jupyter notebook REGRESSION_PY.IPYNB I wrote some Python code for loading the data and performing a linear regression Dengue sequence data. You could use that code I wrote or write one yourself to answer the following numerals. The equivalent R code I wrote can be found in REGRESSION_R.IPYNB, I'm not super versed in R so that code might have a lot of bits to improve.

The data used are genotypes for each of the 4 DENV serotypes and are stored in FASTA (.fas) files. I copy paste this from the linked Wikipedia entrance: A sequence begins with a greater-than character (">") followed by a description of the sequence (all in a single line).

When discussing your results remember to embed the biology of the system in the discussion. Not only limit to the statistical results. And the written answers should be as succinct as possible.

1. In the probabilistic version of linear regression with one variable one want to explain a response variable Y using a single feature X. The main assumption is that the conditional expectation of the measurement is linear in the explanatory variable as shown below. This way one is directly treating in the modeling measurement error ε.

$$\mathbb{E}[Y|X] = \theta_0 + \theta_1 X \tag{1}$$

For one observation one can write the linear regression including the measurement error ε_i. And assume that the measurement errors are independent and have zero mean: $\mathbb{E}[\varepsilon|X] = 0$.

$$Y_i = \theta_0 + \theta_1 x_i + \varepsilon_i \tag{2}$$

1.1 Show that Equations 1 and 2 are equivalent.

1.2 Suppose you know through measurements some statistics of the experiments: $\mathbb{E}[Y]$, $\mathbb{E}[X]$, $\mathrm{Cov}[X,Y]$ and $\mathrm{Var}[Y]$. Write the values of θ_0 and θ_1 in terms of those statistics. You might want to use the properties of conditional expectation listed at the end of the document.

2. Because of the Antibody dependent enhancement (ADE) and lifelong homotypic serotype cross-protection the antigenic distance between sequential pair of infections

of viruses of the same serotype (homotypic infection), should be more consistent than that between sequential pair of infections of viruses of different serotype (heterotypic infection). Furthermore, the antigenic distance between heterotypic pairs should depend on the magnitude of the cross-protection of the first virus against that secondary infecting strain.

The explanatory variable is the Percentage identity distance (PID) and measures the sequence the percentage of matches between two sequences. Thus homotypic pair of viruses have a lower PID, than heterotypic pair of viruses. For the explanatory variable use $x_1 = 1 - \text{PID}$ such that higher values indicate higher similarity between the sequences. However, it is not clear what the antigenic distance between a pair will be because there are substantial selection pressures including but not limited to: ADE, homotypic cross-protection, seasonality in transmission. It is also a virus that has two hosts: mosquitoes and humans, but most of the viral reproduction happens inside the human host.

2.1 **Homotypic infections.** Compare how well a linear regression model fitted to the data corresponding to pair of viruses of the same serotype, e.g. DENV1-DENV1, predict each unseen pair of viruses. For example if your linear model was trained on DENV1-DENV1 pairs, use each other pair: DENV2-DENV2, DENV3-DENV3 and DENV4-DENV4 as the test data.

a. Use the Residual Sum of Squares (RSS) as the goodness-of-fit, and use Figures to present and discuss your results.

b. Use the coefficient of determination (r^2) as the goodness-of-fit, and use Figures to present and discuss your results. The mathematical expression for r^2 is shown below. $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ is the expected value of y.

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{predicted variability}}{\text{observed variability}}$$

2.2 **Heterotypic infections.** Compare how well a linear regression model fitted to the data corresponding to pair of viruses of different serotype, e.g. DENV1-DENV2, predict each unseen pair of viruses. For example if your linear model was trained on DENV1-DENV2 pairs, use each other pair: DENV1-DENV3, DENV1-DENV4, DENV2-DENV3, and DENV3-DENV4 as the test data. Use the Residual Sum of Squares (RSS) as the the goodness-of-fit, use Figures and discuss your results.

a Use the Residual Sum of Squares (RSS) as the goodness-of-fit, and use Figures to present and discuss your results.

b. Use the coefficient of determination (r^2) as the goodness-of-fit, and use Figures to present and discuss your results.

3. Phenotypes have variability across traits and species, such differences are known to emerge from different genetic background and environmental variation. Phenotypic differences on clonal populations (populations of individuals with identical genotypes) exist, and environmental variation is not necessarily responsible for that difference. Thus, some variation observed in the antigenic map could to some extend be a product of this component of phenotypic variance. If inherited this variability could also be a target of selection.

   Assume that the distance of a strain i to the center of it's respective serotype s_i phenotype $D_i^{s_i}$ could represent a measure of how far was strain i from a *mean* strain.

   Use the absolute difference between the distance of a pair of strains i, j to the center of their respective serotypes $|D_i^{s_i} - D_j^{s_j}|$ as another explanatory variable of the linear regression as show below. The first explanatory variable is a measure of nucleotide divergence between a pair of sequences i, j $x_1 = 1 - \text{PID}(i, j)$, and the second explanatory variable is the absolute difference of the distances to the center of each respective serotype $x_2 = |D_i^{s_1} - D_j^{s_2}|$. Don't forget to normalize the explanatory variables such that they have zero mean and unit variance.

   When discussing the results remember that in the ideal world you want to optimize for a model that is both accurate (low Bias), and confident (low variance).

$$Y_i = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

   3.1 How could you visualize correlations between the explanatory variables? Plot it for both homotypic and heterotypic pairs and discuss.

   3.2 Repeat 2.1, discuss your result comparing to the results obtained in numeral 2. Just use r^2 to present and discuss your results.

   3.3 Repeat 2.2, discuss your result comparing to the results obtained in numeral 2. Just use r^2 to present and discuss your results.

   3.4 Use the F statistic to ask if the difference of the distance to the center x_2 is significant in explaining the antigenic distance for each model in 2.1. and 2.2. The F statistic is presented page 48 of the reference, and is also show below, the sub-indexes 0 represent the smaller model while the sub-indexes 1 represent the bigger model, p_i is the number of parameters in model i and n is the number of measurements. Discuss your result comparing to the results obtained in numeral 2.

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(n - p_1 - 1)}$$

   3.5 Given that you're just increasing one explanatory variable what other statistic could you use to investigate if x_2 is significant in explaining the difference in phenotype?

4. In Figure 1 shown below I present the phylogeny that represent the evolutionary history of all DENV serotypes. The method I used for create the tree allows inference of the Time to most recent common ancestor (TMRCA) between a pair of strains among other variables. The x-axis in this tree correspond to a real time and each tip is a sequence.

   If you need help understanding trees I recommend this tutorial: Evolutionary trees: A primer.

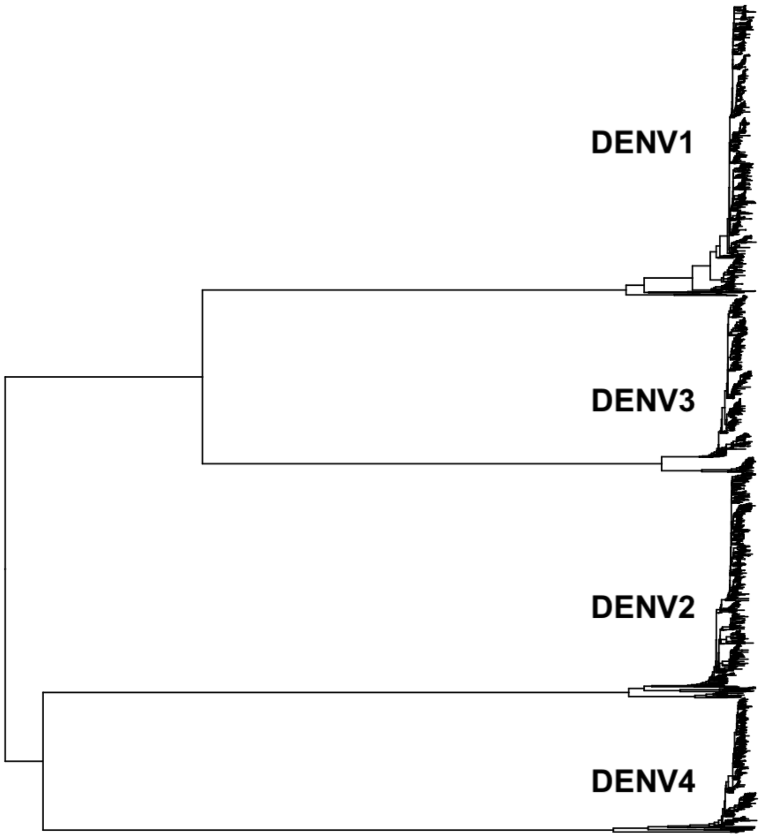

Figure 1: DENV phylogeny

   In the file DATA/DENGUE/TREE/TMRCA.CSV I compiled the TMRCA between all pair of strains calculated from the tree in the file DATA/DENGUE/TREE/TREE.NWK.

   4.1 Use the TMRCA as another explanatory variable. For each serotype in the homotypic and heterotypic pairs visualize the correlation between the explanatory variables and discuss your results.

   4.2 Repeat 3.2 but include a third explanatory variable x_3 with the TMRCA between a pair of strains. Just use r^2 to discuss and present your results.

   4.3 Repeat 3.3 but with the TMRCA between pair of strains, as the in the previous numeral. Just use r^2 to discuss and present your results.

4.4 Use a statistical test to compare your new model to the one proposed in numeral 2.

4.5 Use a statistical test to compare your new model to the one proposed in numeral 3.

Conditional expectation

- If X_1 and X_2 are two random variables and are independent then $\mathbb{E}[X_1|X_2] = \mathbb{E}[X_1]$.
- If a is a constant then $\mathbb{E}[aX_2|X_1] = a\mathbb{E}[X_2|X_1]$. It also works for functions on the conditional variable $a = a(X_1)$.
- The conditional expectation is linear: $\mathbb{E}[X_1 + X_2|X_3] = \mathbb{E}[X_1|X_3] + \mathbb{E}[X_2|X_3]$.
- I've heard this with multiple names, so I'm not sure the *true* one, but I use it as the **Tower property**.

$$\mathbb{E}[\mathbb{E}[X_1|X_2]] = \mathbb{E}[X_1]$$