Exercises - Week 11

Genomics and bioinformatics

1 Transcription Equilibrium

Solve for m the differential equation

$$\dot{m}(t) = P - \gamma \cdot m(t) \quad ,$$

where $m(t)$ is the quantity of RNA in the cell at time t, P is a constant transcription rate, and γ is a constant degradation rate.

2 Quantile Normalization

The expression of a few genes (g1 to g5) has been quantified and reported in the table below, for two technical replicates R1 and R2.

	g1	g2	g3	g4	g5
R1	0.8	1.5	1.7	2.6	3.9
R2	1.1	2.2	1.6	2.6	3.8

1. What is the median of the values in R1, and in R2?
2. Apply quantile normalization to these two samples, by hand.
3. Write a function that applies quantile normalization to any couple of vectors.
4. Test it: create two random vectors of size 2'000, apply quantile normalization to them, and check that after transformation they have the same median.

3 Linear Models

3.1 Continuous variable

The overall gene expression of an individual has been quantified (by a numerologist) at different body temperatures and reported in the table below.

Temperature	-25	-10	-5	0	5	10	25
Expression	13	18	19	22	24	32	37

We want to model the gene expression as a linear function of the temperature. The problem is to find the "best" straight line across our points. A line is defined by its intercept and its slope. Roughly speaking, if we find a slope that is nearly zero, the temperature has no effect on gene expression, while a bigger slope means bigger effect.

1. Plot in R the expression against the temperature. Do you think a linear model can be appropriate? Do you think that the temperature has an effect?

2. Define a linear model of the form $Y = \mathrm{a} + bT + \epsilon$: what does each of the terms correspond to in our problem (with the usual notation, close to the course's) ?

3. Apply the model to our data: each measurement of T and Y is a "realization" of this model, so we get 7 equations:

$$y_1 = a + bt_1 + \epsilon_1 \ ,$$

$$y_2 = a + bt_2 + \epsilon_2 \ ,$$

$$\ldots$$

   Rewrite this system as a single equation in matrix form: $Y = X\beta + \epsilon$, where this time $Y, \epsilon \in \mathbb{R}^7$, $X \in \mathbb{R}^{7\mathrm{x}2}$, and $\beta = \binom{a}{b}$. Replace t_i's and y_i's by their values.

4. We are interested in estimating from our data the intercept a and the slope b, so that the the sum of all square measurement errors (ϵ) is minimal. Let

$$X = (\mathbb{1}, T) = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \ldots & \ldots \end{pmatrix}.$$

   The usual estimator for β is

$$\hat{\beta} = (X'X)^{-1}X'Y,$$

   where X' denotes X transposed.
   Use either R or Python's *numpy* library to compute this vector.
   Hints: In R, the matrix product is `%*%` ; use `solve()` for the inverse.

5. Load your data (T and Y vectors) into R and do the regression the usual way, as follows:

```
model = lm(Y~T)
summary(model)
```

   Visualize the result:

```
plot(T,Y)
abline(model)
```

   - Identify the intercept a, the slope b, the p-value for each (quality of their estimation), the "R-squared" value (overall quality of the model (0 to 1; closer to 1 is better)).
   - Compare with your own estimation of β.
   - Does the temperature significantly affect gene expression?
   - If one increases the body temperature by 1 degree, by how much does it increase the gene expression?
   - Is the linear fit a good choice?

3.2 Categorical variable

The expression of a gene has been quantified several times (replicates) in two individuals and reported in the table below. One individual has been treated by some drug, the other has not. We now want to model the gene expression as a linear function of the binary variable "Treated/Untreated".

	R1	R2	R3	R4	R5
Untreated	41	29	55	50	40
Treated	43	35	60	53	42

Assign (arbitrary) numeric values to the factors: “Treated”=1, “Untreated”=0.

1. Plot in R the expression against the treatment, similarly to the figure in the course’s slides. Do you think that the treatment has an effect?

2. Use the `boxplot` function of R to better compare graphically the two individuals.

3. Perform the same analysis as above and answer the same questions.