Exercises - Week 11

Genomics and bioinformatics

1 Transcription Equilibrium

$\frac{dm(t)}{dt} = P - \gamma m(t) \iff \frac{1}{P-\gamma m(t)}\, dm(t) = dt \iff \int_{m(0)}^{m(t)} \frac{1}{P-\gamma u}\, du = \int_0^t ds$
$\iff \frac{-1}{\gamma}(\log(P-\gamma m(t)) - \log(P-\gamma m(0))) = t - 0 \iff \log(\frac{P-\gamma m(t)}{P-\gamma m(0)}) = -\gamma t$
$\iff \frac{P-\gamma m(t)}{P-\gamma m(0)} = e^{-\gamma t} \iff m(t) = \frac{1}{-\gamma} \cdot (e^{-\gamma t}(P-\gamma m(0)) - P) = -\frac{P}{\gamma}\, e^{-\gamma t} + m(0)\, e^{-\gamma t} + \frac{P}{\gamma}$

which tends to $\frac{P}{\gamma}$ when t tends to infinity ($e^{-\gamma t}$ tends to zero).

2 Quantile Normalization

1. R1: 1.7 - R2: 2.2.

2. Record the initial order in R1 (g1-g2-g3-g4-g5) and R2 (g1-g3-g2-g4-g5). Sort their values:

R1	0.8	1.5	1.7	2.6	3.9
R2	1.1	1.6	2.2	2.6	3.8

   Calculate the mean of each pair:

Avg	0.95	1.55	1.95	2.6	3.85

   Replace R1 and R2 by these same values:

R1'	0.95	1.55	1.95	2.6	3.85
R2'	0.95	1.55	1.95	2.6	3.85

   Reorder as it was initialy:

	g1	g2	g3	g4	g5
R1	0.95	1.55	1.95	2.6	3.85
R2	0.95	1.95	1.55	2.6	3.85

   Now the median and all other quantiles are the same in R1 and R2.

3. 4. 5. : see `week11_solution.R` .

3 Linear Models

3.1 Continuous variable

1. At first view points are roughly aligned and expression seems to increase with temperature. To plot it:

```
T = c(-25,-10,-5,0,5,10,25)
Y = c(13,18,19,22,24,32,37)
plot(T, Y, xlim=c(-30,30), ylim=c(0,50))
```

2. Y is the response: a random variable generating the gene expression values, assumed normally distributed.
   T is the factor: the temperature.
   a is the intercept: the point where the line crosses the vertical axis (unknown).
   b is the slope of the line (unknown).
   ϵ is the measurement error (a normally distributed random variable).

3. $$13 = a - 25b + \epsilon_1 \ ,$$
   $$18 = a - 10b + \epsilon_2 \ ,$$
   $$...$$
   In matrix form:
   $$\begin{pmatrix} 13 \\ 18 \\ 19 \\ 22 \\ 24 \\ 32 \\ 37 \end{pmatrix} = \begin{pmatrix} 1 & -25 \\ 1 & -10 \\ 1 & -5 \\ 1 & 0 \\ 1 & 5 \\ 1 & 10 \\ 1 & 25 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{pmatrix}$$

4.
```
# "%*%" is the matrix product
# solve() gives the inverse
# Vectors are "vertical" by default

X = cbind(rep(1,7),T)
beta = solve(t(X) %*% X ) %*% (t(X) %*% Y)
```
   One finds $a = 23.57$ and $b = 0.51$.

5. The output is the following:

```
> summary(lm(Y~T))

Call:
lm(formula = Y ~ T)

Residuals:
      1       2       3       4       5       6       7
 2.1786 -0.4714 -2.0214 -1.5714 -2.1214  3.3286  0.6786

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.57143    0.88744  26.561 1.41e-06 ***
T            0.51000    0.06062   8.413 0.000389 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 . 0.1 ' ' 1

Residual standard error: 2.348 on 5 degrees of freedom
Multiple R-squared:  0.934,Adjusted R-squared:  0.9208
F-statistic: 70.77 on 1 and 5 DF,  p-value: 0.0003891
```

The Intercept-Estimate is the a, 23.57; the T-Estimate is the b, 0.51. We recognize the numbers that we found with the given formula for β. The "Pr(>|t|)" column contains the p-values. The one for b is low enough to say that the temperature has a significant effect on gene expression. For instance, an increase of 1 in temperature induces an increase of 0.51 in expression. The R-squared is very close to 1, which means that the points lie close to the fitted line, thus the linear model is probably appropriate.

3.2 Categorical variable

1.
```
untreated = c(41,29,55,50,40)
treated = c(43,35,60,53,42)
T = c( rep(0,5), rep(1,5) )
Y = c( untreated, treated )
plot(T, Y, xlim=c(-0.5,1.5), ylim=c(20,70))
```
We notice a systematic increase in the treated sample, but the variance is big and it is hard to decide if there really is an effect.

2.
```
boxplot(Y~T, names=c("Untreated","Treated"))
```

3. The system can be written
$$\begin{aligned} 41 &= a + 0 \cdot b + \epsilon_1 \,, \\ 29 &= a + 0 \cdot b + \epsilon_2 \,, \\ &\cdots \\ 43 &= a + 1 \cdot b + \epsilon_1 \,, \\ 35 &= a + 1 \cdot b + \epsilon_2 \,, \\ &\cdots \end{aligned}$$
or
$$\begin{pmatrix} 41 \\ 29 \\ 55 \\ 50 \\ 40 \\ 43 \\ 35 \\ 60 \\ 53 \\ 42 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \end{pmatrix}$$

4.
```
X = cbind(rep(1,10),T)
beta = solve(t(X) %*% X ) %*% (t(X) %*% Y)
```
One finds $a = 43$ and $b = 3.6$.

5. The output is the following:

```
> summary(lm(Y~T))

Call:
lm(formula = Y ~ T)

Residuals:
   Min     1Q Median     3Q    Max
-14.00  -4.35  -2.50   6.85  13.40

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   43.000      4.447   9.668 1.09e-05 ***
T              3.600      6.290   0.572    0.583
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.945 on 8 degrees of freedom
Multiple R-squared:  0.03934,Adjusted R-squared:  -0.08074
F-statistic: 0.3276 on 1 and 8 DF,  p-value: 0.5828
```

This time the estimate for b has a p-value of about 0.6, which means one cannot trust the result at all. Overall the fit is terrible with an R-squared of 0.04 and a bad F-statistic. However, the estimate for b, 3.6, is the systematic increase that we noticed earlier.