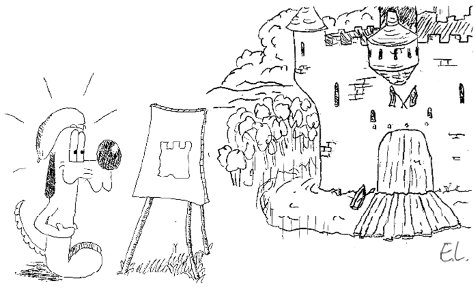
# PERFORMANCE EVALUATION EXERCISES

# MODEL FITTING

Jean-Yves Le Boudec, Spring 2021



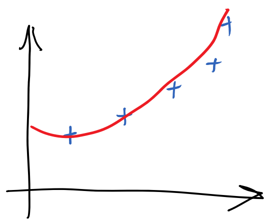
1. We want to fit a data set $y_i$ to a polynomial of degree 2: $y_i \approx at_i^2 + bt_i + c$. Is this a linear regression model ?

   (a) ☐ Yes

   (b) ☐ It depends on the score function

   (c) ☐ It depends on the data set

   (d) ☐ No



2. If the error terms in a fitting model are not homoscedastic, it is better to...

   (a) ☐ Use $\ell^1$-norm minimization rather than $\ell^2$

   (b) ☐ Use weights to make the error term homoscedastic

   (c) ☐ Use $\ell^2$-norm minimization rather than $\ell^1$

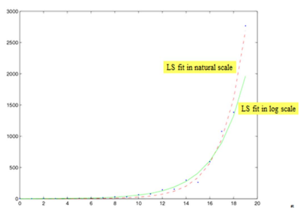
3. The green estimation is least square fit in $y$-log-scale. This corresponds to assuming that the *relative* error terms (blue dot $-$ green curve)/ green curve are...

   (a) ☐ iid

   (b) ☐ approximately normal

   (c) ☐ (a) and (b)

   (d) ☐ None



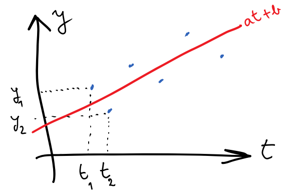

4. We fit the model $Y_i = at_i + b$ using least squares. The obtained line is such that...

(a) ☐ The average vertical distance from the points to the line is 0.

(b) ☐ The average square distance from the points to the line is 0.

(c) ☐ It leaves an equal number of points on each side.

(d) ☐ None of these.



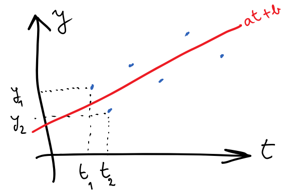

5. We fit the model $Y_i = at_i + b$ using $\ell^1$ norm minimization. The obtained line is such that...

   (a) ☐ The average vertical distance from the points to the line is 0.

   (b) ☐ The average square distance from the points to the line is 0.

   (c) ☐ It leaves an equal number of points on each side.

   (d) ☐ None of these.




6. We have two sets of measurements of the same quantity $\mu$, with different uncertainty. We model this as follows. We have $m+n$ independent measurements $X_1, ..., X_m \sim$ iid $N_{\mu,\sigma^2}$ and $Y_1, ..., Y_n \sim$ iid $N_{\mu,\lambda^2\sigma^2}$. The term $\lambda$ is known.

   (a) What is the maximum likelihood estimate of $\mu$ ?

      i. ☐ $\hat{\mu}_1 = \frac{X_1+...+X_m+Y_1+...Y_n}{m+n}$

      ii. ☐ $\hat{\mu}_2 = \frac{X_1+...+X_m+\lambda(Y_1+...Y_n)}{m+\lambda n}$

      iii. ☐ $\hat{\mu}_3 = \frac{X_1+...+X_m+\frac{Y_1+...Y_n}{\lambda}}{m+\frac{n}{\lambda}}$

      iv. ☐ $\hat{\mu}_4 = \frac{X_1+...+X_m+\frac{Y_1+...Y_n}{\lambda^2}}{m+\frac{n}{\lambda^2}}$

   (b) We now assume that the terms $\sigma$ and $\lambda$ are unknown but we know that $\lambda >> 1$ and $m \approx n$. Which of the following is the best estimate of $\mu$ ?

      i. ☐ $\hat{\mu}_5 = \frac{X_1+...+X_m}{m}$

      ii. ☐ $\hat{\mu}_6 = \frac{Y_1+...Y_n}{n}$

      iii. ☐ $\hat{\mu}_3 = \frac{X_1+...+X_m+Y_1+...Y_n}{m+n}$

   (c) We continue to assume that $\lambda >> 1$ and $m \approx n$, furthermore we assume that the value of $\lambda$ is known (but $\sigma$ is not known). Which of the following is the best estimate of $\mu$ ?

      i. ☐ $\hat{\mu}_5 = \frac{X_1+...+X_m}{m}$

      ii. ☐ $\hat{\mu}_6 = \frac{Y_1+...Y_n}{n}$

      iii. ☐ $\hat{\mu}_3 = \frac{X_1+...+X_m+Y_1+...Y_n}{m+n}$

iv. $\square$ $\hat{\mu}_4 = \frac{X_1+...+X_m+\frac{Y_1+...Y_n}{\lambda^2}}{m+\frac{n}{\lambda^2}}$

7. We consider the following model for the virus infection data example:

$$\log Y_i = \log a + \alpha t_i + \varepsilon_i \text{ with } \varepsilon_i \sim \text{ iid } N_{0,\sigma^2},\ ,i = 1...I$$

The goal of this exercise is to apply Theorem 3.3 to this model. To simplify the notation, let $L_i = \log Y_i$ and $\ell = \log a$, so that the model is

$$L_i = \ell + \alpha t_i + \varepsilon_i \text{ with } \varepsilon_i \sim \text{ iid } N_{0,\sigma^2}\ ,i = 1...I$$

It will also be convenient to use $\bar{t} = \frac{1}{I}\sum_{i=1}^{I} t_i$, $\bar{L} = \frac{1}{I}\sum_{i=1}^{I} L_i$, $v = \frac{1}{I}\sum_{i=1}^{I}(t_i - \bar{t})^2$ (sample variance of $t$) and $c = \frac{1}{I}\sum_{i=1}^{I}\left(L_i t_i - \bar{t}\bar{L}\right)$ (sample covariance of $t$ and $\log Y$).

(a) Write the matrix $X$.

(b) Does assumption (H) hold ?

(c) Compute $X^T X$ as a function of $\bar{t}$ and $v$ and verify that it is invertible when $H$ holds.

(d) Let $\vec{L} = \begin{pmatrix} L_1 \\ ... \\ L_I \end{pmatrix}$. Compute $X^T \vec{L}$ as a function of $\bar{t}$, $\bar{L}$ and $c$. Compute $G = (X^T X)^{-1}$

(e) Let $s^2$ be the rescaled sum of squared residuals (you are not asked to compute $s$). Give the formulae, derived from Theorem 3.3, for 95% confidence intervals for $\ell$ and for $\alpha$.