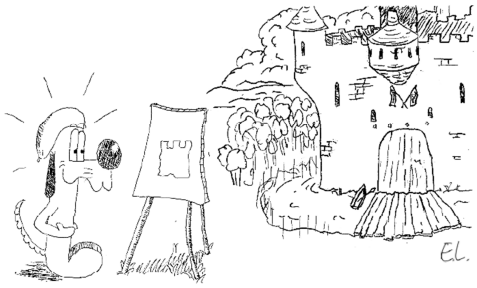
Performance Evaluation Exercises

Model Fitting

With Solutions Jean-Yves Le Boudec, Spring 2021



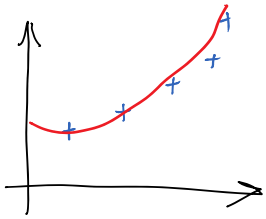
1. We want to fit a data set y_i to a polynomial of degree 2: $y_i \approx at_i^2 + bt_i + c$. Is this a linear regression model ?

   (a) ☒ Yes

   (b) ☐ It depends on the score function

   (c) ☐ It depends on the data set

   (d) ☐ No



Solution. Linear regression means linear with respect to the parameters of the model other than the noise terms. Here the parameters are a, b, c and the model depends linearly on them.

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

   (a) ☐ Use ℓ^1-norm minimization rather than ℓ^2

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

   (c) ☐ Use ℓ^2-norm minimization rather than ℓ^1

Solution. If the error terms are not homoscedastic, it is better to change the model, for example by rescaling, or to use weights in the ℓ^1 or ℓ^2 scores. Changing from ℓ^1 to ℓ^2 or vice-versa does not solve the problem.

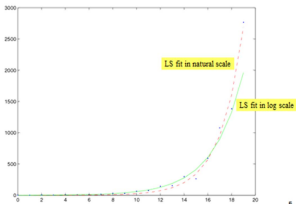
3. The green estimation is least square fit in y-log-scale. This corresponds to assuming that the *relative* error terms (blue dot − green curve)/ green curve are...

   (a) ☐ iid

   (b) ☐ approximately normal

   (c) ☒ (a) and (b)

   (d) ☐ None



Solution. The relative error terms are

$$\frac{Y_i - f_i(\beta)}{f_i(\beta)} = \frac{f_i(\beta)e^{\varepsilon_i} - f_i(\beta)}{f_i(\beta)} = e^{\varepsilon_i} - 1$$

They are independent and identically distributed.

They are the exponential of a normal random variable, thus they are not normal (such distributions are called a shifted log-normal distribution).

If the errors are small, then

$$e^{\varepsilon_i} = 1 + \varepsilon_i + o(\varepsilon_i) - 1 = \varepsilon_i + o(\varepsilon_i) \approx \varepsilon_i$$

and the error terms are approximately iid normal (i.e. homoscedastic).

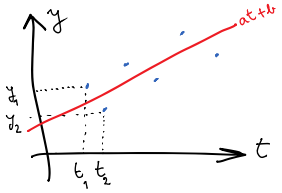
4. We fit the model $Y_i = at_i + b$ using least squares. The obtained line is such that...

   (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.



Solution. Let $\hat{a}, \hat{b}$ the fitted parameters. $\hat{a}, \hat{b}$ minimizes

$$\sum_i (y_i - (at_i + b))^2 \text{ over } a, b \in \mathbb{R}$$

Thus $\hat{b}$ minimizes

$$\sum_i (y_i - (\hat{a}t_i + b))^2 \text{ over } b \in \mathbb{R}$$

Let $x_i \stackrel{\text{def}}{=} y_i - \hat{a}t_i$. We have: $\hat{b}$ minimizes

$$\sum_i (x_i - b)^2 \text{ over } b \in \mathbb{R}$$

We saw in class that the solution to this problem is the average $\hat{b} = \bar{x} = \frac{1}{I} x_i$ (I is the number of points). Thus the average of the vertical distances to the lines, $x_i - b$, is

$$\frac{1}{I} \sum_{i=1}^{I} \left(x_i - \hat{b} \right) = \bar{x} - \hat{b} = 0$$

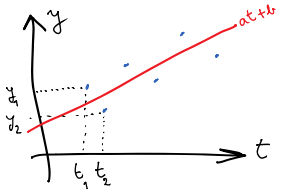
5. We fit the model $Y_i = at_i + b$ using ℓ^1 norm minimization. The obtained line is such that...

   (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.



Solution. Let $\hat{a}, \hat{b}$ the fitted parameters. $\hat{a}, \hat{b}$ minimizes

$$\sum_i |y_i - (at_i + b)| \text{ over } a, b \in \mathbb{R}$$

Thus $\hat{b}$ minimizes

$$\sum_i |y_i - (\hat{a}t_i + b)| \text{ over } b \in \mathbb{R}$$

Let $x_i \stackrel{\text{def}}{=} y_i - \hat{a}t_i$, i.e. x_i is the vertical distance to the line with offset 0. We have: $\hat{b}$ minimizes

$$\sum_i |x_i - b| \text{ over } b \in \mathbb{R}$$

We saw in class that the solutions to this problem are the medians of x. Thus $\hat{b}$ is a median of x, namely at least half of the values of x_i are $\leq \hat{b}$ and at least half of them are $\geq \hat{b}$. Thus at least half of the values of the vertical distances $x_i - \hat{b}$ are ≤ 0 and at least half of them are ≥ 0.

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

   (a) What is the maximum likelihood estimate of μ ?

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

   Solution. It is a case of weighted least square fitting. The MLE minimizes the score defined by

   $$\sum_i (x_i - \mu)^2 + \sum_i \frac{(y_i - \mu)^2}{\lambda^2}$$

   The derivative with respect to μ is

   $$-2\sum_i (x_i - \mu) - 2\sum_j \frac{y_j - \mu}{\lambda^2}$$

   The optimal is

   $$\begin{aligned} \hat{\mu} &= \frac{x_1 + ... + x_m + \frac{y_1+...y_n}{\lambda^2}}{m + \frac{n}{\lambda^2}} \\ &= \frac{m\bar{x} + \frac{n\bar{y}}{\lambda^2}}{m + \frac{n}{\lambda^2}} \end{aligned}$$

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

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

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

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

Solution. None of these estimates is the maximum likelihood estimator, seen in the previous question. Since we can only choose from these 3, let us see how good they are. First note that all 3 estimators have expectation equal to μ, so we can compare them by comparing their variances.

$$\begin{aligned}
\mathrm{var}(\hat{\mu}_5) &= \frac{\sigma^2}{m} \\
\mathrm{var}(\hat{\mu}_6) &= \frac{\lambda^2\sigma^2}{n} >> \mathrm{var}(\hat{\mu}_1) \\
\mathrm{var}(\hat{\mu}_3) &= \frac{\sigma^2\left(m+\lambda^2 n\right)}{(m+n)^2} \approx \frac{\sigma^2}{m}\frac{(1+\lambda^2)}{4} >> \mathrm{var}(\hat{\mu}_1)
\end{aligned}$$

In the given conditions, (λ large and m, n comparable) it is better to forget the measurements with large variance than to average all measurements. But if n is much larger than n this may not hold.

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

i. ☐ $\hat{\mu}_5 = \frac{X_1+\ldots+X_m}{m}$
ii. ☐ $\hat{\mu}_6 = \frac{Y_1+\ldots Y_n}{n}$
iii. ☐ $\hat{\mu}_3 = \frac{X_1+\ldots+X_m+Y_1+\ldots Y_n}{m+n}$
iv. ☒ $\hat{\mu}_4 = \frac{X_1+\ldots+X_m+\frac{Y_1+\ldots Y_n}{\lambda^2}}{m+\frac{n}{\lambda^2}}$

Solution. Since $\hat{\mu}_5$ is better than $\hat{\mu}_6$ and $\hat{\mu}_3$, the only remaining question is whether $\hat{\mu}_4$ is better than $\hat{\mu}_5$. The expectation of $\hat{\mu}_4$ is μ like for the other 3 and its variance is

$$\mathrm{var}(\hat{\mu}_4) = \frac{\sigma^2}{m+\frac{n}{\lambda^2}} < \mathrm{var}(\hat{\mu}_5) = \frac{\sigma^2}{m}$$

thus $\hat{\mu}_4$ is the best estimator.

This is not surprising as it is the maximum likelihood estimator and such estimators are asymptotically optimal. With a bit of more work, it can be shown that the variance of $\hat{\mu}_4$ is the smallest of all 4 estimators for all values of $m, n, \lambda^2, \sigma^2$ (even when the conditions $m \approx n$ and $\lambda >> 1$ do not hold).

Take-home message: adding some very noisy measurements may lead to worst estimation, unless you compensate for the increased noise by appropriately weighting the estimator !

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

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

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

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

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^TX as a function of $\bar{t}$ and v and verify that it is invertible when H holds.

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

(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 ℓ and for α.

Solution. Consider that the parameter is $\beta = \begin{pmatrix} \ell \\ \alpha \end{pmatrix}$. Then

(a)
$$X = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ & \dots \\ 1 & t_I \end{pmatrix}$$

(b) Here $p = 2$ and (H) means that X is of rank 2. This holds if at least two rows of X are not colinear, i.e. the t_i's are not all equal.

(c)
$$X^TX = \begin{pmatrix} I & \sum_i t_i \\ \sum_i t_i & \sum_i t_i^2 \end{pmatrix} = I \begin{pmatrix} 1 & \bar{t} \\ \bar{t} & v + \bar{t}^2 \end{pmatrix}$$
where we used the equality $\frac{1}{I}\sum_i t_i^2 = v + \bar{t}^2$.
X^TX is invertible if and only if its determinant is non zero. Its determinant is
$$I^2 \begin{vmatrix} 1 & \bar{t} \\ \bar{t} & v + \bar{t}^2 \end{vmatrix} = I^2 v$$
and is non zero if and only if the sample variance v of t is non zero, i.e. when all t_is are not equal – this is precisely assumption (H), as expected.

(d)
$$X^T\vec{L} = \begin{pmatrix} \sum_i L_i \\ \sum_i t_i L_i \end{pmatrix} = I \begin{pmatrix} \bar{L} \\ c + \bar{t}\bar{L} \end{pmatrix}$$
where we used the equality $\frac{1}{I}\sum_i t_i L_i = c + \bar{t}\bar{L}$.
The inverse of the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, when $ad - bc \neq 0$, is $\frac{1}{ad-bc}\begin{pmatrix} d & -c \\ -b & a \end{pmatrix}$, therefore:
$$G = \frac{1}{Iv}\begin{pmatrix} v + \bar{t}^2 & -\bar{t} \\ -\bar{t} & 1 \end{pmatrix}$$
By Theorem 3.3, the maximum likelihood estimate is
$$\begin{pmatrix} \hat{\ell} \\ \hat{\alpha} \end{pmatrix} = (X^TX)^{-1}X^T\vec{L} = GX^T\vec{L} = \frac{1}{Iv}\begin{pmatrix} v + \bar{t}^2 & -\bar{t} \\ -\bar{t} & 1 \end{pmatrix} I \begin{pmatrix} \bar{L} \\ c + \bar{t}\bar{L} \end{pmatrix}$$
$$= \frac{1}{v}\begin{pmatrix} v\hat{L} - c\bar{t} \\ c \end{pmatrix} = \begin{pmatrix} \hat{L} - \frac{c}{v}\bar{t} \\ \frac{c}{v} \end{pmatrix}$$
i.e.
$$\begin{aligned} \hat{\ell} &= \bar{L} - \hat{\alpha}\bar{t} \\ \hat{\alpha} &= \frac{c}{v} \end{aligned}$$
These are the classical least-square formulae for fitting a straight line to a 2d-cloud of points. Note the simple formula for the slope α.

(e) Apply item 4 of the theorem. We obtain a confidence interval for ℓ by setting $u_1 = 1$, $u_2 = 0$. The variance bias for $\hat{\ell}$ is

$$g = G_{1,1} = \frac{1}{I}\left(1 + \frac{\bar{t}^2}{v}\right)$$

Let η be the 97.5% quantile of the Student-$(I-2)$ distribution (when I is large, $\eta \approx 1.96$). A 95% confidence interval for ℓ is

$$\hat{\ell} \pm \eta s\sqrt{g} = \hat{\ell} \pm \eta s \frac{\sqrt{1 + \frac{\bar{t}^2}{v}}}{\sqrt{I}}$$

For $\hat{\alpha}$, the corresponding term is obtained by setting $u_1 = 0$, $u_2 = 1$ and is

$$g = G_{2,2} = \frac{1}{Iv}$$

A 95% confidence interval for α is

$$\hat{\alpha} \pm \eta s\sqrt{g} = \hat{\alpha} \pm \eta s \frac{1}{\sqrt{Iv}}$$