Problem formulation

A nonlinear least-squares problem is an optimization problem on the form
\[
\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} \sum_{i=1}^{m} r_i(x)^2,
\]
where \( n \) is the number of variables.

The objective function \( f(x) \) is defined by \( m \) auxiliary residual functions \{\( r_i(x) \)\}.

We will assume that \( m \geq n \).

The problem is called least-squares since we are minimizing the sum of squares of the residual functions.

If we assume that the development of the population is exponential, the model function might be
\[
g(t) = x_1 e^{x_2 t}
\]
and the residuals
\[
r_i(x) = g(t_i) - y_i = x_1 e^{x_2 t_i} - y_i.
\]

In standard least squares problems, the vertical distance (squared) between observations and a model function are minimized.
We will write the optimization problem as
\[
\min_x f(x),
\]
where
\[
f(x) = \frac{1}{2} \sum_{i=1}^{m} \eta_i(x)^2 \equiv \frac{1}{2} r(x)^T r(x) \equiv \frac{1}{2} \|r(x)\|^2,
\]
and \(r\) is a vector-valued function
\[
r(x) = [r_1(x) \ r_2(x) \ \ldots \ r_m(x)]^T.
\]
For each value of \(x\), the residual function value \(r(x)\) may be interpreted as a point in “observation space” \(\mathbb{R}^m\).

The residual function describes a (usually \(n\)-dimensional) surface in \(\mathbb{R}^m\).

Observe that
\[
\min_x \frac{1}{2} \|r(x)\|^2
\]
may be interpreted as
\[
\min_x \frac{1}{2} \|r(x) - 0\|^2.
\]

Thus, a least squares problem may be interpreted as trying to find the point \(x^*\) in parameter space \(\mathbb{R}^n\) that corresponds to the point \(r(x^*)\) in observation space \(\mathbb{R}^m\) that is closest to the origin.

For the antelope data and model
\[
f(x) = \frac{1}{2} \sum_{i=1}^{5} (x_1 e^{\alpha t_i} - y_i)^2 = \frac{1}{2} r(x)^T r(x),
\]
\[
\begin{bmatrix}
x_1 e^{\alpha t_1} - y_1 \\
x_1 e^{\alpha t_2} - y_2 \\
x_1 e^{\alpha t_3} - y_3 \\
x_1 e^{\alpha t_4} - y_4 \\
x_1 e^{\alpha t_5} - y_5
\end{bmatrix}
= \begin{bmatrix}x_1 e^{1\alpha} - 3 \\
x_1 e^{2\alpha} - 4 \\
x_1 e^{3\alpha} - 6 \\
x_1 e^{4\alpha} - 11 \\
x_1 e^{5\alpha} - 20
\end{bmatrix}.
\]

Then,
\[
\min_x \frac{1}{2} \|r(x)\|^2 = \min_x \frac{1}{2} \|h(x) - y\|^2
\]

With this formulation we are trying to find the point \(x^*\) in parameter space \(\mathbb{R}^n\) that corresponds to the point \(h(x^*)\) on the model surface in observation space \(\mathbb{R}^m\) that is closest to the observations \(y\).
\[ r(x) = [r_1(x) \ r_2(x) \ r_3(x)]^T \text{ residual surface in } \mathbb{R}^m \]

\[ h(x) = [h_1(x) \ h_2(x) \ h_3(x)]^T \text{ observation surface in } \mathbb{R}^m \]
\( r(x) = [r_1(x) \ r_2(x) \ r_3(x)]^T \) residual surface in \( \mathbb{R}^m \)

\[
f(x) = \frac{1}{2} \| r(x) \|^2 \text{ as a function of } x = [x_1 \\ x_2]^T \text{ in } \mathbb{R}^n
\]
Observation space $\mathbb{R}^3$

Parameter space $\mathbb{R}^2$

Model space

- $h_1(t) = 2.53e^{0.21t}$
- $h_2(t) = 3.57e^{0.10t}$
- $h_3(t) = 5.29e^{0.44t}$
Observation space $\mathbb{R}^3$

Parameter space $\mathbb{R}^2$

Model space

$\begin{align*}
h_1(t) &= 2.53e^{0.21t} \\
h_2(t) &= 3.57e^{0.10t} \\
h_3(t) &= 5.29e^{0.44t}
\end{align*}$
Parameter space $\mathbb{R}^2$

Observation space $\mathbb{R}^3$

Model space

\[ h_1(t) = 2.53e^{0.21t} \]
\[ h_2(t) = 3.57e^{0.10t} \]
\[ h_3(t) = 5.29e^{0.44t} \]
The Gradient, Jacobian, and Hessian

The Gradient and Hessian structure

\[ \nabla f(x) = \frac{1}{2} \left( J(x)^T J(x) + \sum_{i=1}^{m} r_i(x)^2 r_i(x) \right), \]

where \( J(x) \) is the Jacobian of \( r(x) \), i.e.

\[ \nabla f(x) = \nabla r(x)^T J(x), \]

and may be derived from the chain rule.

For the example data and model

\[ \nabla f(x) = (j(x)^T j(x)) + \sum_{i=1}^{m} r_i(x)^2 r_i(x), \]

the Hessian is

\[ \nabla^2 f(x) = \nabla r(x)^T J(x). \]

Thus, the Hessian of a least-squares objective function is a sum of two terms:

- \( j(x)^T j(x) \) with first-order derivatives only, and
- \( \sum_{i=1}^{m} r_i(x)^2 r_i(x) \) with second-order derivatives.

The Gauss-Newton method

A method that uses the approximation \( Q(x) = 0 \) is called the

\[ \nabla f(x) = \nabla r(x)^T J(x), \]

or with the Hessian approximated by \( (j(x)^T j(x))^{-1} j(x)^T \), i.e.

\[ \nabla f(x) = \nabla r(x)^T J(x). \]

Using the chain rule again on

\[ \nabla f(x) = \nabla r(x)^T J(x), \]

\[ \nabla^2 f(x) = \nabla r(x)^T J(x). \]
The Gauss-Newton method

The Newton formulation

- If we assume that $J(x)$ has full rank, the Hessian approximation
  \[ J(x)^T J(x) \]
  is positive definite and the Gauss-Newton search direction $p^{GN}$ is a descent direction.
- Otherwise, $J(x)^T J(x)$ is non-invertible and the equation
  \[ J(x)^T J(x)p^{GN} = -J(x)^T r(x) \]
  does not have a unique solution. In this case, the problem is said to be under-determined or over-parameterized.

The Gauss-Newton method

Geometrical interpretation of the search direction

- The linear approximation corresponds to a tangent plane to the surface $r(x)$ at $r_k = r(x_k)$.
- The point on the tangent plane closest to the origin is given by the projection of $-r_k$ onto the range space of $J_k$, since
  \[ J_k p^{GN} = J_k (J_k^T J_k)^{-1} J_k^T (-r_k). \]

The Gauss-Newton method

The linear least squares formulation

- Assume we approximate the residual function $r(x)$ with a linear Taylor function, i.e. the plane
  \[ r(x_k + p) \approx r(x_k) + J_k p. \]
- The minimizer on the plane is found by solving the linear least squares problem
  \[ \min_p \frac{1}{2} \| J_k p + r_k \|^2 = \min_p \frac{1}{2} \| J_k p - (-r_k) \|^2. \]
- The solution is given by the normal equations
  \[ J_k^T J_k p = -J_k^T r_k \]
  or
  \[ p = (J_k^T J_k)^{-1} J_k^T (-r_k). \]
- Thus, the minimizer on the plane corresponds to the Gauss-Newton search direction.

The Gauss-Newton method

Geometrical interpretation of the search direction

- The first order condition
  \[ \nabla f(x^*) = 0 \]
  corresponds to when
  \[ J(x^*)^T r(x^*) = 0, \]
  i.e. when $r(x^*)$ is orthogonal to the tangent plane spanned by the columns of $J(x^*)$. 
**Geometric interpretation of the first order condition**

**Zero residual problems**

- A special case is when \( r(x^*) = 0 \Rightarrow f(x^*) = 0 \).
- In this case the problem is said to have zero residual and the surface \( r(x) \) intersects the origin.

**Convergence for the Gauss-Newton method**

- If \( r(x^*) = 0 \), the approximation \( Q(x) \approx 0 \) is good and the Gauss-Newton method will behave like the Newton method close to the solution, i.e. converge quadratically if \( J(x^*) \) has full rank.
- The advantage over the Newton method is that we do not need to calculate the second-order derivatives \( \nabla^2 r_i(x) \).
- However, if any residual component \( r_i(x^*) \) and/or the corresponding curvature \( \nabla^2 r_i(x) \) is large, the approximation \( Q(x) \approx 0 \) will be poor, and the Gauss-Newton method will converge slower than the Newton method.
- For such problems, the Gauss-Newton method may not even be locally convergent, i.e. without a global strategy such as the line search, it will not converge no matter how close to the solution we start.

**Statistical interpretation**

**Stochastic model**

- If the residuals are interpreted statistically, i.e. we have a model
  \[
  y_i = x_1 e^{x_2 t} + \varepsilon_i
  \]
  and the errors \( \varepsilon_i \) are assumed to be independent and normally distributed \( N(0, \sigma^2) \), our least squares estimation of the parameters will be the maximum likelihood estimators given our measurement \( y_i \).

**Variance of estimated parameters**

- The variance for the estimated parameters are calculated from the variance-covariance matrix
  \[
  D = \sigma^2 (\nabla^2 f(x^*))^{-1},
  \]
  where each diagonal element \( d_{ii} \) correspond to the variance of the parameter \( x_i \), and the off-diagonal element \( d_{ij} \) correspond to the covariance between parameters \( x_i \) and \( x_j \).
- If \( \sigma^2 \) is unknown, it may be estimated by
  \[
  \hat{\sigma}^2 = \frac{r(x^*)^T r(x^*)}{m - n},
  \]
  where \( m \) is the number of observations, and \( n \) is the number of parameters.
Statistical interpretation

Variance of estimated parameters

- A high variance means a high degree of uncertainty about a parameter.
- In this context, the inverse matrix
  \[ K = D^{-1} = \frac{1}{\sigma^2} \nabla^2 f(x^*) , \]
  is sometimes called the information matrix, since higher diagonal values \( k_i \) correspond to more information about the parameter \( x_i \).
- Since the information matrix is proportional to the hessian \( H(x^*) = \nabla^2 f(x^*) \), strong curvature corresponds to high information, i.e. good localization of the parameter.
- Furthermore, if the hessian is approximated by
  \[ \nabla^2 f(x^*) \approx J(x^*)^T J(x^*) , \]
  the covariances will only be first order approximations of the true covariances.

Statistical interpretation

Variance of estimated parameters

- The approximations of the parameters and the covariances makes it possible to derive confidence limits, do hypothesis testing, etc.
- For linear problems, the covariance estimations are exact. For non-linear problems, the covariances are still exact, but the confidence limits are not, since the confidence regions are not ellipses.

Weighted least squares

Formulation

- If the observations errors are dependent and/or with different variances, weighted least squares should be used, i.e. the problem
  \[ \min_x r(x)^T W r(x) , \]
  should be solved.
- If the matrix \( \Sigma \) with elements \( \sigma_{ij}^2 \) contain the covariances between observations \( i \) and \( j \), the optimal choice of \( W \) is
  \[ W = \Sigma^{-1} , \]
  and the solution of the weighted least squares problem is again the maximum likelihood solution.
- The distance measure \( r(x)^T \Sigma^{-1} r(x) \) is sometimes called the Mahalanobis distance.
If the observations are independent, $\Sigma$ and $W$ will be diagonal matrices, and $w_i = 1/\sigma_i^2$.

Thus the solution will rely more on “good” observations, since residuals with a corresponding small observation error will be weighted more heavily in the objective function.

**Orthogonal regression**

- When we solve the problem
  \[
  \min_x \frac{1}{2} \sum_{i=1}^m r_i(x)^2 = \frac{1}{2} \min_x r(x)^T r(x)
  \]
  where $r_i(x) = g(t_i) - y_i$ is the difference between our model and our measured values, we minimize the square of the vertical distance.

  If we assume that we have errors also in the independent variable $t_i$, it may be appropriate to minimize the orthogonal distance between the model and the measurements instead.

  This may be formulated as solving the problem
  \[
  \min_{x, \delta} f(x) = \sum_{i=1}^m r_i(x; t_i + \delta_i)^2 + \|\delta\|^2,
  \]
  where $\delta_i$ is the error in $t_i$ and $r_i(x; t_i + \delta_i) = g_i(t_i + \delta_i) - y_i$.

  Problem minimizing the orthogonal distance between model and measurements are sometimes referred to as orthogonal regression problems.

**Weighted least squares**

**Methods**

- If we want to solve a weighted least squares problem, there are two equivalent solutions:
  - Modify the algorithm or
  - modify the residual/Jacobian function.

- A modified algorithm would solve the following equation
  \[
  J^T W J p = -J^T W r.
  \]

- A modified residual/Jacobian would be
  \[
  r_s(x) = R r(x), J_s(x) = R J(x),
  \]
  where $R^T R = W$ is the Cholesky factorization of $W$.

- Such a factor $R$ will always exist if $W$ is positive semidefinite.

**Orthogonal regression**

**Methods**

- By reformulating the objective function, we may use algorithms for “conventional” non-linear least squares to solve orthogonal regression problems.

  For our example
  \[
  y = g(t) = x_1 e^{x_2 t}
  \]
  we may introduce one point $(s_i, g(s_i))$ on the curve for each measurement $(t_i, y_i)$.

- Defining the component function $r_i(x)$
  \[
  r_i(x) = \begin{bmatrix} g(t_i) - y_i \\ s_i - t_i \\ \vdots \\ r_m(x) \end{bmatrix},
  \]
  and $r(x) = \begin{bmatrix} r_1(x) \\ \vdots \\ r_m(x) \end{bmatrix}$,

  the least squares problem takes the following, standard, form:
  \[
  \min_x r(x)^T r(x).
  \]