Notes on Linear Regression

Paul F. Roysdon, Ph.D.

I. Introduction to Linear Regression

Consider the linear equation from high school algebra

$$y = mx + b$$

where $y, m, x, b \in \mathbb{R}^1$, all scalars. Let the parameters m, b define the slope and y-intercept, respectively Assume that the values of parameters m, b are given. The value x is the input, and y is the output. This simple equation takes a single input to produce a single output – a point on a 2D plane. If multiple inputs are used, producing multiple outputs, the result are points along a line on a 2D plane. By computing the output for each given input, using the parameters m, b, the user is producing a line is space.

Now consider the case where *data* is provided, e.g. (x_i, y_i) for $i = 1, \ldots, m$, not the *parameters* that define the data (i.e. the variables m, b for the example above). The "reverse-engineering" process to use the data to find the parameters is called regression, or estimation, because we are finding "regressors" for the parameters or "estimating" parameters.

The most basic, and certainly the most common, method is Least Squares (LS). LS assumes no statistical properties of the data, i.e. if the data has some noise, and the mean and variance of the noise is known, the LS method does not account for noise and "blindly" finds the best fit, minimizing the squared error of the residuals.

However, we can do one step better, while still avoiding any statistical assumptions. If we can find a suitable "weighting" matrix, to apply weights to the data, a Weighted Least Squares (WLS) can be solved. While the algorithm formulation is general, the guess-and-check method is still required to determine the "optimal" weights.

Now assume that we know, or can assume, the statistical properties of the data, and we know the *mean* and *variance* of the data, then we can formulate WLS into the Maximum Likelihood Estimator (MLE). Generally speaking, in this class of estimators, the MLE is the optimal estimator.

Example: given vectors $\mathbf{a} = 1980, 1990, 2000, 2010$ and $\mathbf{y} = 94, 184, 195, 192$ we wish to fit a line to the data. Our first test will be a linear "model" for $\mathbf{H}$. In a linear model there are two parameters to estimate, the slope and the y-intercept. Because the slope is dependent on the input x we need to estimate the slope as a *variable*, whereas the y-intercept is not dependent on x and thus we estimate it as a *constant*. Mathematically our model will have a column for the variables, and a column for the constant

$$\mathbf{H} = \begin{bmatrix} a_1 & 1 \\ a_2 & 1 \\ a_3 & 1 \\ a_4 & 1 \end{bmatrix}$$

using the linear model

$$\mathbf{y} = \mathbf{H}\boldsymbol{x}$$

solve

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} a_1 & 1 \\ a_2 & 1 \\ a_3 & 1 \\ a_4 & 1 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix}$$

$$\begin{bmatrix} 94 \\ 184 \\ 195 \\ 192 \end{bmatrix} = \begin{bmatrix} 1980 & 1 \\ 1990 & 1 \\ 2000 & 1 \\ 2010 & 1 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix}$$

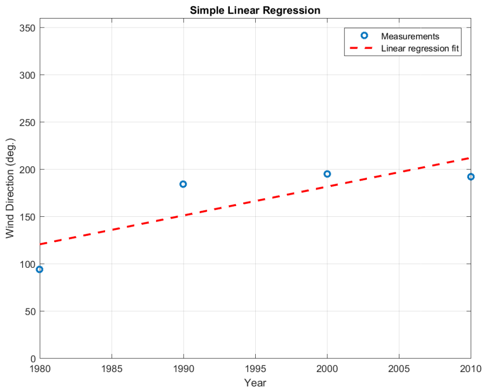
solving for the estimator

$$\hat{\boldsymbol{x}} = (\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{y}$$

we find the values for $\hat{\boldsymbol{x}}$. Now compute the points along the line using our "estimated" parameters, where the equation of the line (in vector form)

$$\hat{y}_i = \hat{x}_1 a_i + \hat{x}_2$$

for $i = 1, \ldots, 4$. Notice here that we are computing the estimated vector $\hat{\mathbf{y}}$ using our estimated parameters in $\hat{\boldsymbol{x}}$.



Extending the linear model to a quadratic model, with parameter vector $\boldsymbol{\beta}$, estimated variables $\hat{\boldsymbol{x}}$, and output estimate $\hat{y}$, then

$$\hat{y} = \beta_2^2 \hat{x}_2 + \beta_1 \hat{x}_1 + \beta_0 \hat{x}_0$$

and

$$\mathbf{H} = \begin{bmatrix} a_1^2 & a_1 & 1 \\ a_2^2 & a_2 & 1 \\ a_3^2 & a_3 & 1 \\ a_4^2 & a_4 & 1 \end{bmatrix}.$$

The solution to the new model is computed using the same steps presented above.

II. Least Squares

Consider the general measurement equation

$$\mathbf{y} = \mathbf{H}\boldsymbol{x} + \boldsymbol{\eta} + \mathbf{e}$$

where $\mathbf{y} \in \mathbb{R}^{m\times 1}$, $\mathbf{H} \in \mathbb{R}^{m\times n}$ where $m > n$ and $\text{rank}(\mathbf{H}) = n$, $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$, with

Gaussian noise $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I}) \in \mathbb{R}^{n\times 1}$,

and deterministic errors $\mathbf{e} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I}) \in \mathbb{R}^{n\times 1}$.

Ignoring the noise and error vectors, the estimate of $\boldsymbol{x}$ is found by

$$\begin{aligned} \mathbf{J}_{LS}(\hat{\boldsymbol{x}}) &= \frac{1}{2}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{x}})^\intercal(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{x}}) \\ &= \frac{1}{2}(\mathbf{y}^\intercal\mathbf{y} - 2\mathbf{y}^\intercal\mathbf{H}\hat{\boldsymbol{x}} + \boldsymbol{x}^\intercal\mathbf{H}^\intercal\mathbf{H}\hat{\boldsymbol{x}}) \\ \frac{\partial \mathbf{J}_{LS}(\hat{\boldsymbol{x}})}{\partial \hat{\boldsymbol{x}}} &= -\mathbf{H}^\intercal\mathbf{y} + \mathbf{H}^\intercal\mathbf{H}\hat{\boldsymbol{x}} = 0 \\ \hat{\boldsymbol{x}} &= (\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{y} \\ &= \bar{\mathbf{H}}\mathbf{y} \end{aligned}$$

where $\bar{\mathbf{H}} \triangleq (\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal$ is the generalized inverse, also known as the “Moore-Penrose pseudo-inverse”. Note that $\bar{\mathbf{H}}$ transforms the measurement space to the state space. If $\mathbf{H}$ is full column-rank, then $\mathbf{H}$ has the following property

$$\bar{\mathbf{H}}\mathbf{H} = (\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{H} = \mathbf{I}_m$$

This is because $\mathbf{H}^\intercal\mathbf{H} \in \mathbb{R}^{m\times m}$ with $\text{rank}(\mathbf{H}^\intercal\mathbf{H}) = m$, and therefore nonsingular. Then by the linear algebra property for the general matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$ with $\text{rank}(\mathbf{A}) = m$, the property $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_m$ is applied in eqn. (1).

By analysis, the estimate $\hat{\boldsymbol{x}}$ is

$$\begin{aligned} \hat{\boldsymbol{x}} &= \bar{\mathbf{H}}\mathbf{y} \\ &= \bar{\mathbf{H}}(\mathbf{H}\boldsymbol{x} + \boldsymbol{\eta} + \mathbf{e}) \end{aligned}$$

The estimation error is

$$\begin{aligned} \delta\boldsymbol{x} &= \boldsymbol{x} - \hat{\boldsymbol{x}} \\ &= \boldsymbol{x} - \bar{\mathbf{H}}(\mathbf{H}\boldsymbol{x} + \boldsymbol{\eta} + \mathbf{e}) \\ &= -\bar{\mathbf{H}}(\boldsymbol{\eta} + \mathbf{e}) \end{aligned}$$

The measurement estimate is

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{H}\hat{\boldsymbol{x}} \\ &= \mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{y} \\ &= \mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal(\mathbf{H}\boldsymbol{x} + \boldsymbol{\eta} + \mathbf{e}) \\ &= \mathbf{P}\mathbf{H}\boldsymbol{x} + \mathbf{P}(\boldsymbol{\eta} + \mathbf{e}) \\ &= \mathbf{H}\boldsymbol{x} + \mathbf{P}(\boldsymbol{\eta} + \mathbf{e}) \end{aligned}$$

where the projection matrix $\mathbf{P} \triangleq \mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal$.

The measurement residual is

$$\begin{aligned} \mathbf{r} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= (\mathbf{H}\boldsymbol{x} + \boldsymbol{\eta} + \mathbf{e}) - \mathbf{H}\boldsymbol{x} - \mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal(\boldsymbol{\eta} + \mathbf{e}) \\ &= (\mathbf{I}_m - \mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal)(\boldsymbol{\eta} + \mathbf{e}) \\ &= (\mathbf{I}_m - \mathbf{P})(\boldsymbol{\eta} + \mathbf{e}) \\ &= \mathbf{Q}(\boldsymbol{\eta} + \mathbf{e}) \\ &= \mathbf{Q}\boldsymbol{\eta} + \mathbf{Q}\mathbf{e} \end{aligned}$$

where the orthogonal projection matrix $\mathbf{Q} \triangleq (\mathbf{I}_m - \mathbf{P})$.

Projection matrices $\mathbf{P}$ and $\mathbf{Q}$ are both idempotent, and have rank n and $m - n$ respectively. The proofs for idempotent and rank are presented in Section IV.

III. Weighted Least Squares

Consider the general measurement equation

$$\mathbf{y} = \mathbf{H}\boldsymbol{x} + \boldsymbol{\nu}$$

where $\mathbf{y} \in \mathbb{R}^{m\times 1}$, $\mathbf{H} \in \mathbb{R}^{m\times n}$ where $m > n$ and $\text{rank}(\mathbf{H}) = n$, $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$, with Gaussian noise $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I}) \in \mathbb{R}^{n\times 1}$.

Ignoring the noise, the estimate of $\boldsymbol{x}$ is found by

$$\begin{aligned} \mathbf{J}_{WLS}(\hat{\boldsymbol{x}}) &= \frac{1}{2}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{x}})^\intercal\mathbf{W}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{x}}) \\ &= \frac{1}{2}(\mathbf{y}^\intercal\mathbf{W}\mathbf{y} - 2\mathbf{y}^\intercal\mathbf{W}\mathbf{H}\hat{\boldsymbol{x}} + \boldsymbol{x}^\intercal\mathbf{H}^\intercal\mathbf{W}\mathbf{H}\hat{\boldsymbol{x}}) \\ \frac{\partial \mathbf{J}_{WLS}(\hat{\boldsymbol{x}})}{\partial \hat{\boldsymbol{x}}} &= -\mathbf{H}^\intercal\mathbf{W}\mathbf{y} + \mathbf{H}^\intercal\mathbf{W}\mathbf{H}\hat{\boldsymbol{x}} = 0 \\ \hat{\boldsymbol{x}} &= (\mathbf{H}^\intercal\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{W}\mathbf{y} \end{aligned}$$

where $\mathbf{W} \in \mathbb{R}^{m\times m}$ is the weighting matrix.

The estimation error is

$$\begin{aligned} \delta\boldsymbol{x} &= \boldsymbol{x} - \hat{\boldsymbol{x}} \\ &= \boldsymbol{x} - (\mathbf{H}^\intercal\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{W}\mathbf{y} \\ &= \boldsymbol{x} - (\mathbf{H}^\intercal\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{W}(\mathbf{H}\boldsymbol{x} + \boldsymbol{\nu}) \\ &= \big(\mathbf{I} - (\mathbf{H}^\intercal\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{W}\mathbf{H}\big)\,\boldsymbol{x} \\ &\quad - (\mathbf{H}^\intercal\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{W}\boldsymbol{\nu} \\ &= -(\mathbf{H}^\intercal\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{W}\boldsymbol{\nu} \end{aligned}$$

For $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I})$, $\mathbf{W} = \mathbf{R}^{-1}$

$$\begin{aligned} \text{E}\left\langle \delta\boldsymbol{x} \right\rangle &= \mathbf{0} \\ \text{var}\left\langle \delta\boldsymbol{x} \right\rangle &= (\mathbf{H}^\intercal\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{W}\mathbf{R}\mathbf{W}\mathbf{H}(\mathbf{H}^\intercal\mathbf{W}\mathbf{H})^{-1} \end{aligned}$$

For $\mathbf{W} = \mathbf{I}_m$, the Least Squares (LS) estimate results

$$\begin{aligned} \hat{\boldsymbol{x}} &= (\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{y} \\ \text{E}\left\langle \delta\boldsymbol{x} \right\rangle &= \mathbf{0} \\ \text{var}\left\langle \delta\boldsymbol{x} \right\rangle &= (\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{R}\mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1} \end{aligned}$$

For $\mathbf{W} = \mathbf{R}^{-1}$, the Maximum Likelihood Estimate (MLE) results

$$\begin{aligned}
\hat{\boldsymbol{x}} &= (\mathbf{H}^\intercal\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{R}^{-1}\mathbf{y} \\
\mathrm{E}\langle\delta\boldsymbol{x}\rangle &= \mathbf{0} \\
\mathrm{var}\langle\delta\boldsymbol{x}\rangle &= (\mathbf{H}^\intercal\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{R}^{-1}\mathbf{R}\mathbf{R}^{-1}\mathbf{H}(\mathbf{H}^\intercal\mathbf{R}^{-1}\mathbf{H})^{-1} \\
&= (\mathbf{H}^\intercal\mathbf{R}^{-1}\mathbf{H})^{-1} \\
&= \mathbf{C}
\end{aligned}$$

where $\mathbf{C}$ is the covariance matrix, and $\mathbf{C}^{-1} = \mathbf{H}^\intercal\mathbf{R}^{-1}\mathbf{H}$ is the information matrix.

IV. Proof of Matrix Rank Using the SVD

A. *Proof of idempotent P*

For the matrix $\mathbf{P}$ to be idempotent, it must be the case that $\mathbf{P} = \mathbf{P}^\intercal\mathbf{P} = \mathbf{P}\mathbf{P}$, where $\mathbf{P} \triangleq \mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal$, and $\mathbf{H} \in \mathbb{R}^{m\times n}$, with $m > n$. Thus we can show:

$$\begin{aligned}
\mathbf{P}^\intercal &= (\mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal)^\intercal \\
&= \mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal \\
&= \mathbf{P} \\
\mathbf{P}\mathbf{P} &= \mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal \\
&= \mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal \\
&= \mathbf{P} \\
\therefore \mathbf{P}^\intercal\mathbf{P} &= \mathbf{P}\mathbf{P} = \mathbf{P}.
\end{aligned}$$

■

B. *Proof of rank P*

We can prove that rank$(\mathbf{P}) = n$. First recall that $\mathbf{H} \in \mathbb{R}^{m\times n}$, with $m > n$ and full column rank, i.e. rank$(\mathbf{H}) = n$. Let the SVD of $\mathbf{H}$ be defined as

$$\begin{aligned}
\mathbf{H} &= \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\intercal \\
&= [\mathbf{U}_1, \mathbf{U}_2]\left[\begin{pmatrix}\boldsymbol{\Sigma}_1 \\ \boldsymbol{\Sigma}_0\end{pmatrix}\right]\mathbf{V}^\intercal \qquad (1)
\end{aligned}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m}$, $\boldsymbol{\Sigma}_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}$, and $\boldsymbol{\Sigma}_0 = \mathbf{0} \in \mathbb{R}^{(m-n)\times n}$, where σ_i for $i = 1, \ldots, n$ are the singular values of $\mathbf{H}$. Both $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are unitary matrices, therefore $\mathbf{U}\mathbf{U}^\intercal = \mathbf{U}^\intercal\mathbf{U} = \mathbf{I} \in \mathbb{R}^{m\times m}$ and $\mathbf{V}\mathbf{V}^\intercal = \mathbf{V}^\intercal\mathbf{V} = \mathbf{I} \in \mathbb{R}^{m\times m}$. The columns of $\mathbf{U}_1 \in \mathbb{R}^{m\times n}$ form an orthonormal basis for the range-space of $\mathbf{H}$, and the columns of $\mathbf{U}_2 \in \mathbb{R}^{m\times(m-n)}$ form the null-space of $\mathbf{H}^\intercal$. Similarly the first n columns of $\mathbf{V}$ form an orthonormal basis for the range of $\mathbf{H}^\intercal$, and the $m-n$ columns of $\mathbf{V}$ form an orthonormal basis for the null-space of $\mathbf{H}$. Finally, the eigenvectors $\mathbf{V}$ of the matrix $\mathbf{H}^\intercal\mathbf{H}$ are the right singular values of $\mathbf{H}$, and the singular values of $\mathbf{H}$ squared are the corresponding nonzero eigenvalues: $\sigma_i = \sqrt{\lambda_i(\mathbf{H}^\intercal\mathbf{H})}$. Similarly, the eigenvectors of $\mathbf{H}\mathbf{H}^\intercal$ are the left singular vectors $\mathbf{U}$ of matrix $\mathbf{H}$, and the singular values of $\mathbf{H}$ squared are the nonzero eigenvalues of $\mathbf{H}\mathbf{H}^\intercal$: $\sigma_i = \sqrt{\lambda_i(\mathbf{H}\mathbf{H}^\intercal)}$.

Define $\mathbf{P}$ in terms of the SVD of $\mathbf{H}$:

$$\begin{aligned}
\mathbf{P} &= \mathbf{H}(\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal \\
&= (\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\intercal)(\mathbf{V}\boldsymbol{\Sigma}^\intercal\mathbf{U}^\intercal\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^t)^{-1}(\mathbf{V}\boldsymbol{\Sigma}^\intercal\mathbf{U}^\intercal) \\
&= (\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\intercal)(\mathbf{V}\boldsymbol{\Sigma}^\intercal\boldsymbol{\Sigma}\mathbf{V}^\intercal)^{-1}(\mathbf{V}\boldsymbol{\Sigma}^\intercal\mathbf{U}^\intercal) \\
&= (\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\intercal)(\mathbf{V}\boldsymbol{\Sigma}_1^2\mathbf{V}^\intercal)^{-1}(\mathbf{V}\boldsymbol{\Sigma}^\intercal\mathbf{U}^\intercal) \qquad (2) \\
&= (\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\intercal)(\mathbf{V})^{-1}(\boldsymbol{\Sigma}_1^2)^{-1}(\mathbf{V}^\intercal)^{-1}(\mathbf{V}\boldsymbol{\Sigma}^\intercal\mathbf{U}^\intercal) \qquad (3) \\
&= \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\intercal\mathbf{V}\boldsymbol{\Sigma}_1^{-2}\mathbf{V}^\intercal\mathbf{V}\boldsymbol{\Sigma}^\intercal\mathbf{U}^\intercal \qquad (4) \\
&= \mathbf{U}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^{-2}\boldsymbol{\Sigma}_1^\intercal\mathbf{U}^\intercal \qquad (5) \\
&= \mathbf{U}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_1^\intercal\mathbf{U}^\intercal \qquad (6) \\
&= \mathbf{U}\mathbf{I}_{n\times n}\mathbf{U}^\intercal \\
&= [\mathbf{U}_1\mathbf{U}_2]\begin{bmatrix}\mathbf{I}_{n\times n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{U}_1^\intercal \\ \mathbf{U}_2^\intercal\end{bmatrix} \\
&= \mathbf{U}_1\mathbf{U}_1^\intercal.
\end{aligned}$$

The middle product in eqn. (2) can be separated because it is an $n \times n$ matrix with rank n, and it is non-singular. In eqns. (2)-(6), we need only consider $\boldsymbol{\Sigma}_1$ as $\boldsymbol{\Sigma}_0$ drops out.

The rank of matrix $\mathbf{P}$ is defined as the number of non-zero singular values of $\mathbf{P}$. Thus, rank$(\mathbf{P}) = n$. Similarly, because $\mathbf{P}$ is idempotent, rank$(\mathbf{P}) = tr(\mathbf{P})$, then rank$(\mathbf{P}) = n$.

■

C. *Proof of idempotent Q*

For the matrix $\mathbf{Q}$ to be idempotent, it must be the case that $\mathbf{Q} = \mathbf{Q}^\intercal\mathbf{Q} = \mathbf{Q}\mathbf{Q}$, where $\mathbf{Q} \triangleq (\mathbf{I} - \mathbf{P})$, and $\mathbf{P} \in \mathbb{R}^{m\times m}$. Thus we can show:

$$\begin{aligned}
\mathbf{Q}\mathbf{Q} &= (\mathbf{I} - \mathbf{P})(\mathbf{I} - \mathbf{P}) \\
&= \mathbf{I} - \mathbf{P} \\
&= \mathbf{Q} \\
\mathbf{Q}^\intercal\mathbf{Q} &= (\mathbf{I} - \mathbf{P})^\intercal(\mathbf{I} - \mathbf{P}) \\
&= (\mathbf{I} - \mathbf{P}^\intercal)(\mathbf{I} - \mathbf{P}) \\
&= \mathbf{I} - \mathbf{P} - \mathbf{P}^\intercal + \mathbf{P}^\intercal\mathbf{P}, \quad \mathbf{P} = \mathbf{P}^\intercal\mathbf{P} \\
&= \mathbf{I} - \mathbf{P} - \mathbf{P}^\intercal + \mathbf{P}, \quad \mathbf{P}^\intercal = \mathbf{P} \\
&= \mathbf{I} - \mathbf{P} \\
&= \mathbf{Q} \\
\therefore \mathbf{Q}^\intercal\mathbf{Q} &= \mathbf{Q}\mathbf{Q} = \mathbf{Q}
\end{aligned}$$

■

D. *Proof of rank Q*

We can prove that rank$(\mathbf{Q}) = m - n$ by the SVD of $\mathbf{H}$. Apply the result from the proof for the rank of $\mathbf{P}$, where $\mathbf{P} \in \mathbb{R}^{m\times m}$ and $\mathbf{I} \in \mathbb{R}^{m\times m}$. Using the inner product we can define $\mathbf{I}$ in terms of $\mathbf{U}$

$$\begin{aligned}
\mathbf{I} &= \mathbf{U}\mathbf{U}^\intercal \\
&= [\mathbf{U}_1\mathbf{U}_2]\begin{bmatrix}\mathbf{U}_1^\intercal \\ \mathbf{U}_2^\intercal\end{bmatrix} \\
&= \mathbf{U}_1\mathbf{U}_1^\intercal + \mathbf{U}_2\mathbf{U}_2^\intercal.
\end{aligned}$$

Alternatively, by the outer product we can define

$$\begin{aligned}\mathbf{I} &= \mathbf{U}^\intercal \mathbf{U} \\ &= \left[\begin{array}{c}\mathbf{U}_1^\intercal \\ \mathbf{U}_2^\intercal\end{array}\right][\mathbf{U}_1\mathbf{U}_2] \\ &= \left[\begin{array}{cc}\mathbf{U}_1^\intercal\mathbf{U}_1 & \mathbf{U}_1^\intercal\mathbf{U}_2 \\ \mathbf{U}_2^\intercal\mathbf{U}_2 & \mathbf{U}_2^\intercal\mathbf{U}_2\end{array}\right]\end{aligned}$$

where $\mathbf{U}_1^\intercal\mathbf{U}_1 = \mathbf{I} \in \mathbb{R}^{n\times n}$, $\mathbf{U}_2^\intercal\mathbf{U}_2 = \mathbf{I} \in \mathbb{R}^{(m-n)\times(m-n)}$. Finally, $\mathbf{U}_2\mathbf{U}_2^\intercal = \mathbf{P} \in \mathbb{R}^{m\times m}$ as proved above, and $\mathbf{U}_1\mathbf{U}_1^\intercal = \mathbf{Q} \in \mathbb{R}^{m\times m}$ which is proven below.

Now define $\mathbf{Q}$ as

$$\begin{aligned}\mathbf{Q} &= \mathbf{I} - \mathbf{P} \\ &= (\mathbf{U}_1\mathbf{U}_1^\intercal + \mathbf{U}_2\mathbf{U}_2^\intercal) - \mathbf{U}_1\mathbf{U}_1^\intercal \\ &= \mathbf{U}_2\mathbf{U}_2^\intercal.\end{aligned}$$

The rank of matrix $\mathbf{Q}$ is defined as the number of non-zero singular values of $\mathbf{Q}$. Thus, for $\mathbf{Q} \triangleq (\mathbf{I}-\mathbf{P})$, and rank$(\mathbf{P}) = n$, the number of non-zero singular values of $\mathbf{Q}$ is at most $m-n$, and therefore the rank$(\mathbf{Q}) = m-n$.

■

E. *Physical Interpretation of P & Q*

The physical interpretation for $\mathbf{P}$ and $\mathbf{Q}$ is a mapping of the measurement and the residual, as shown in Fig. 1. $\mathbf{Py}$ projects $\mathbf{y}$ onto the range$(\mathbf{P})$ along the direction of $\mathbf{y}$. The complementary projector is $\mathbf{Q}$, where $\mathbf{Qy}$ projects $\mathbf{y}$ onto the range$(\mathbf{Q})$ which is orthogonal to the range$(\mathbf{P})$.

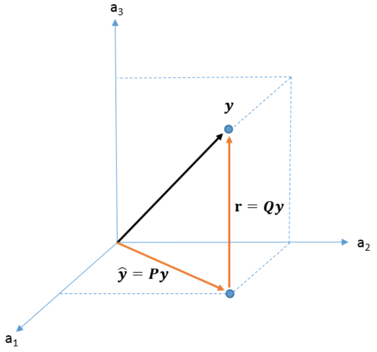


Fig. 1. For a general space in $\mathbb{R}^3$, the mapping $\mathbf{Py} = \hat{\mathbf{y}}$ is the estimate for $\mathbf{y}$, and $\mathbf{Qy} = \mathbf{r}$ is the estimation residual for $\mathbf{y}$.

From the SVD of $\mathbf{H}$ we have the relations:

1) $\mathbf{V}_1\mathbf{V}_1^\intercal$ is the orthogonal projector onto $[N(\mathbf{H})]^\perp = R(\mathbf{H}^\intercal)$.
2) $\mathbf{V}_2\mathbf{V}_2^\intercal$ is the orthogonal projector onto $N(\mathbf{H})$.
3) $\mathbf{U}_1\mathbf{U}_1^\intercal$ is the orthogonal projector onto $R(\mathbf{H})$.
4) $\mathbf{U}_2\mathbf{U}_2^\intercal$ is the orthogonal projector onto $[R(\mathbf{H})]^\perp = N(\mathbf{H}^\intercal)$.