4.4 LEAST SQUARES

The following problem arises in almost all areas where mathematics is applied. At discrete points x_i (e.g., points in time), observations y_i of an event are made, and the results are recorded as a set of ordered pairs

$$\mathcal{D} = \{(x_1, y_1),\ \ (x_2, y_2),\ \ \ldots,\ (x_m, y_m)\} \quad (m > 2). \tag{4.4.1}$$

On the basis of these observations, the goal is to make estimations or predictions at points that are between or beyond the observation made at x_i. The problem boils down finding the equation of a curve $y = f(x)$ that closely fits the points in $\mathcal{D}$ so that the phenomenon can be estimated at any non-observation point $\widetilde{x}$ with the value $\hat{y} = f(\widetilde{x})$.

Traditional (or Ordinary) Least Squares

When the data in $\mathcal{D}$ suggests a linear trend, the traditional theory revolves around the fundamental problem of fitting a straight line to the points in $\mathcal{D}$.

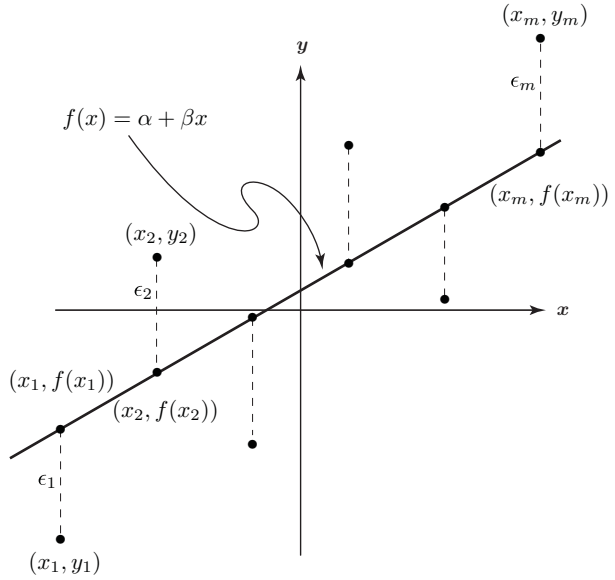


FIGURE 4.4.1: LEAST SQUARES LINE

The strategy is to determine the coefficients α and β in the equation of the line $f(x) = \alpha + \beta x$ that best fits the points (x_i, y_i) in the sense that the sum of the squares of the vertical[†] errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_m$ indicated in Figure 4.4.1 is minimal.

[†] Only vertical errors are considered because there is a tacit assumption that only the observations y_i are subject to error or variation. The x_i's are assumed to be errorless—think of them as being exact points in time, which they often are. If the x_i's are also subject to variation, then horizontal as well as vertical errors in Figure 4.4.1 need to be considered, and a more general theory known as *total least squares* would emerge. The least squares line $\mathcal{L}$ obtained by minimizing only vertical deviations will not be the closest line to points in $\mathcal{D}$

The vertical distance from an observation (x_i, y_i) to a line $f(x) = \alpha + \beta x$ is

$$\epsilon_i = y_i - f(x_i) = y_i - (\alpha + \beta x_i). \tag{4.4.2}$$

Some of the ϵ_i's will be positive while others are negative, so instead of minimizing $\sum_i \epsilon_i$, the aim is to find values for α and β such that

$$\sum_{i=1}^{m} \epsilon_i^2 = \sum_{i=1}^{m} (y_i - \alpha - \beta x_i)^2 \quad \text{is minimal.} \tag{4.4.3}$$

This is the traditional (ordinary) least squares problem. The difference between this and statistical linear regression is that regression considers the y_i's and ϵ_i's to be random variables such that $E[\epsilon_i] = 0$ for each i—i.e., regression includes the hypothesis that the errors "average out to zero." [†] You need not be concerned with the distinction at this point—linear regression is taken up on page 486.

Minimization techniques from calculus say that the minimum value in (4.4.3) must occur at a solution to the system of the two equations

$$\begin{aligned} 0 &= \frac{\partial \left(\sum_{i=1}^{m} (y_i - \alpha - \beta x_i)^2 \right)}{\partial \alpha} = -2 \sum_{i=1}^{m} (y_i - \alpha - \beta x_i), \\ 0 &= \frac{\partial \left(\sum_{i=1}^{m} (y_i - \alpha - \beta x_i)^2 \right)}{\partial \beta} = -2 \sum_{i=1}^{m} (y_i - \alpha - \beta x_i)\, x_i. \end{aligned}$$

Rearranging terms produces two equations in the two unknowns α and β

$$\begin{aligned} \left(\sum_{i=1}^{m} 1 \right) \alpha + \left(\sum_{i=1}^{m} x_i \right) \beta &= \sum_{i=1}^{m} y_i, \\ \left(\sum_{i=1}^{m} x_i \right) \alpha + \left(\sum_{i=1}^{m} x_i^2 \right) \beta &= \sum_{i=1}^{m} x_i y_i. \end{aligned} \tag{4.4.4}$$

By setting $\mathbf{A} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix}$, $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$, and $\widehat{\mathbf{x}} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$, it is seen that the two equations (4.4.4) take the matrix form $\mathbf{A}^T \mathbf{A} \widehat{\mathbf{x}} = \mathbf{A}^T \mathbf{y}$, which is called the

in terms of perpendicular distance, but $\mathcal{L}$ is nevertheless optimal in a certain sense—see the Gauss–Markov theorem on page 491.

[†] The terminology and statistical development of regression analysis was popularized by the English statistician Sir Francis Galton (1822–1911) in his 1886 publication of *Regression Towards Mediocrity in Hereditary Stature* in which he observed that extreme characteristics such as heights of taller and shorter parents are not completely passed on to their children, but rather the characteristics of their children tend to revert or "regress" towards a mediocre point (the mean of all children). Galton was a cousin of Charles Darwin whose book *Origin of Species* stimulated Galton's interest in exploring variation in human populations.

system of normal equations. The normal equations are always consistent (even when $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{y}$ is inconsistent) because $\mathbf{A}^T\mathbf{y} \in R\left(\mathbf{A}^T\right) = R\left(\mathbf{A}^T\mathbf{A}\right)$.

The solution $\widehat{\mathbf{x}}$ of the normal equations $\mathbf{A}^T\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^T\mathbf{y}$ is called a *least squares solution* for the associated system $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{y}$ (generally inconsistent) because $\widehat{\mathbf{x}} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ contains the coefficients in $f(x) = \alpha + \beta x$ (the least squares line) that provides the least squares fit. The vector $\widehat{\mathbf{y}} = \mathbf{A}\widehat{\mathbf{x}}$ is the predicted or estimated vector because its entries $\widehat{y}_i = f(x_i)$ are the least squares estimates of y_i that are predicted by the least squares line. The ϵ_i's in (4.4.2) are the entries in the *residual* (or *error*) vector $\boldsymbol{\epsilon} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}$, so

$$\sum_{i=1} \epsilon_i^2 = \sum (y_i - \widehat{y}_i)^2 = \|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2 = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}})^T(\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}). \tag{4.4.5}$$

This number is referred to as the *error sum of squares*, and it is denoted by SSE.

In the perfect (or ideal) situation when all data points (x_i, y_i) exactly lie on $\mathcal{L}$, then $\boldsymbol{\epsilon} = \mathbf{0}$, and $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{y}$ is a consistent system. But if not all data points are on a straight line, then $\|\boldsymbol{\epsilon}\|_2 > 0$, which in turn means that $\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}} \neq \mathbf{0}$, so that $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{y}$ represents an inconsistent system. This observation can be helpful in identifying the matrix $\mathbf{A}$ involved in setting up more general least squares problems by asking yourself, "what system is required to model an ideal situation?" If $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{y}$ models the "ideal" situation that is not actually realized, then the solution of the associated normal equations $\mathbf{A}^T\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^T\mathbf{y}$ provides the least squares solution.

Partitioning Sums of Squares

In addition to the error sum of squares (SSE) in (4.4.5), there are two other relevant sums of squares. Let $\mu = \mu_{\mathbf{y}} = (\sum_i y_i)/m$ and $\mathbf{e}$ be a column of 1's, and define

SST: *The total sum of squares* $= \sum_{i=1}^m (y_i - \mu)^2 = \|\mathbf{y} - \mu\mathbf{e}\|_2^2$,

SSR: *The regression sum of squares* $= \sum_{i=1}^m (\widehat{y}_i - \mu)^2 = \|\widehat{\mathbf{y}} - \mu\mathbf{e}\|_2^2$.

The relation between SST, SSE, and SSR is revealed in the following theorem.

4.4.1. Theorem. For a set $\{(x_i, y_i)\}$ of $m > 2$ non-colinear data points and for $\mathbf{A} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$, let $\widehat{\mathbf{y}} = \mathbf{A}\widehat{\mathbf{x}}$, where $\widehat{\mathbf{x}}$ is the least squares solution obtained from $\mathbf{A}^T\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^T\mathbf{y}$. It is always true that $\mu_{\widehat{\mathbf{y}}} = \mu = \mu_{\mathbf{y}}$, and

$$\text{SST} = \text{SSE} + \text{SSR}. \tag{4.4.6}$$

Proof. To see that $\mu_{\hat{\mathbf{y}}} = \mu_{\mathbf{y}}$, it suffices to show $\mathbf{e}^T\widehat{\mathbf{y}} = \mathbf{e}^T\mathbf{y}$. The non-colinearity assumption forces $rank\,(\mathbf{A}) = 2$ so that $\mathbf{A}^T\mathbf{A}$ is nonsingular, and hence the solution of $\mathbf{A}^T\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^T\mathbf{y}$ is

$$\widehat{\mathbf{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \implies \widehat{\mathbf{y}} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} = \mathbf{P}\mathbf{y}, \tag{4.4.7}$$

where $\mathbf{P} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{P}^T$ is the orthogonal projector onto $R\,(\mathbf{A})$ (see (4.3.18) on page 461). Furthermore, $\mathbf{e} \in R\,(\mathbf{A})$ means that $\mathbf{P}\mathbf{e} = \mathbf{e}$ so that

$$\mathbf{e}^T\widehat{\mathbf{y}} = \mathbf{e}^T\mathbf{P}\mathbf{y} = \mathbf{e}^T\mathbf{P}^T\mathbf{y} = \mathbf{e}^T\mathbf{y} \implies \mu_{\hat{\mathbf{y}}} = \mu_{\mathbf{y}}.$$

To prove that SST = SSE+SSR, simply verify that $(\mathbf{y}-\widehat{\mathbf{y}}) \perp (\widehat{\mathbf{y}}-\mu\mathbf{e})$ (Exercise 4.4.2) and invoke the Pythagorean theorem (page 40) to conclude that

$$\|\mathbf{y} - \mu\mathbf{e}\|_2^2 = \|(\mathbf{y} - \widehat{\mathbf{y}}) + (\widehat{\mathbf{y}} - \mu\mathbf{e})\|_2^2 = \|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2 + \|\widehat{\mathbf{y}} - \mu\mathbf{e}\|_2^2. \quad \blacksquare$$

Coefficient of Determination

The significance of being able to partition the total sum of squares as indicated in (4.4.6) can now be understood. The sample correlation coefficient between the observed vector $\mathbf{y}$ and the predicted vector $\widehat{\mathbf{y}} = \mathbf{A}\widehat{\mathbf{x}}$ is

$$r = r_{\mathbf{y}\hat{\mathbf{y}}} = \frac{\|\widehat{\mathbf{y}} - \mu\mathbf{e}\|_2}{\|\mathbf{y} - \mu\mathbf{e}\|_2} = \frac{s_{\hat{\mathbf{y}}}}{s_{\mathbf{y}}} \qquad \text{(see (1.6.10) on page 46).}$$

While r may be of some interest, it is not as important as r^2, which is given a special name.

4.4.2. Definition. The term

$$r^2 = r^2_{\mathbf{y}\hat{\mathbf{y}}} = \frac{\|\widehat{\mathbf{y}} - \mu\mathbf{e}\|_2^2}{\|\mathbf{y} - \mu\mathbf{e}\|_2^2} = \frac{s^2_{\hat{\mathbf{y}}}}{s^2_{\mathbf{y}}} = \frac{\mathrm{Var}[\widehat{\mathbf{y}}]}{\mathrm{Var}[\mathbf{y}]} = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}. \tag{4.4.8}$$

is called the ***coefficient of determination.***

The utility of r^2 stems from consideration of variation. The total variation $\mathrm{SST} = \|\mathbf{y} - \mu\mathbf{e}\|_2^2 = \sum(y_i - \mu_{\mathbf{y}})^2$ is the variation in $\mathbf{y}$ from its mean without regard to variations in $\mathbf{x}$. But if, for example, $\mathbf{y}$ tends to vary linearly with $\mathbf{x}$ in a positive manner, then the y_i's generally increase as the x_i's increase, so a more important issue is, "how much (or what percentage) of the total variation in $\mathbf{y}$ is explained by the variation in $\mathbf{x}$ as determined by the least squares line $\mathcal{L}$?" The term $\mathrm{SSE} = \|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2 = \sum(y_i - \widehat{y}_i)^2$ is the variation in $\mathbf{y}$ that is *not* explained by $\mathcal{L}$ (i.e., by the variation in $\mathbf{x}$). Since $\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$, the proportion (or percentage) of the total variation in $\mathbf{y}$ that *is* explained by $\mathcal{L}$ is

$$1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = \frac{\mathrm{SSR}}{\mathrm{SST}} = r^2.$$

Goodness of Fit

A primary use of the coefficient of determination $0 \leq r^2 \leq 1$ is to assess how well the least squares line $\mathcal{L}$ fits the data (x_i, y_i). Each data point is exactly on $\mathcal{L}$ if and only if $\text{SSE} = \|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2 = 0$—i.e, if and only if $r^2 = 1$. This means that *all* of the variation in $\mathbf{y}$ is completely explained by $\mathcal{L}$. At the other extreme, $r^2 = 0$ if and only if $\text{SSR} = \|\widehat{\mathbf{y}} - \mu\mathbf{e}\|_2^2 = 0$, which means that *none* of the variation in $\mathbf{y}$ is explained by $\mathcal{L}$. This is equivalent to saying that $\mathcal{L}$ is perfectly horizontal and that there is not a linear relationship between $\mathbf{x}$ and $\mathbf{y}$. For example if $r^2 = .85$, then 85% of the variation of $\mathbf{y}$ is explained by $\mathcal{L}$. Whether or not this translates to saying that the $\mathcal{L}$ is a good fit for the data can be subjective and application dependent, but it nevertheless provides more insight than looking only at the raw residual $\text{SSE} = \sum_{i=1} \epsilon_i^2 = \sum (y_i - \widehat{y}_i)^2 = \|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2$.

Example (Sales Estimation)

Suppose that a company has been in business for four years, and the sales y_i for year x_i (in tens of thousands of dollars) is shown in the table in Figure 4.4.2. Plotting the data points (x_i, y_i) for $x_1 = 1, x_2 = 2, x_3 = 3$, and $x_4 = 4$ indicates that they do not exactly lie on a straight line, but nevertheless there is a linear trend in sales. Consequently, to predict the sales for a future year it is reasonable to fit the linear trend with a straight line $f(x) = \alpha + \beta x$ that best fits the data in the sense of least squares.

Year x_i	Sales y_i
1	23
2	27
3	30
4	34

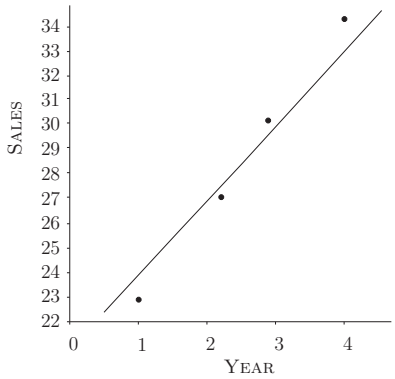


Figure 4.4.2: Linear sales trend

If sales were exactly linear, then there would exist an α and β such that $y_i = \alpha + \beta x_i$ for each $i = 1, 2, 3, 4$ so that $\begin{pmatrix} 23 \\ 27 \\ 30 \\ 34 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$, or equivalently, $\mathbf{y} = \mathbf{A}\widehat{\mathbf{x}}$. But sales are not exactly linear, so $\epsilon_i = y_i - (\alpha + \beta x_i) \neq 0$ for at least one i, or equivalently, $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{A}\widehat{\mathbf{x}} \neq \mathbf{0}$. Least squares theory guarantees that the solution $\widehat{\mathbf{x}}$ of the associated system of normal equations

$$\mathbf{A}^T\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^T\mathbf{y} \implies \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} 114 \\ 303 \end{pmatrix} \implies \widehat{\mathbf{x}} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} 19.5 \\ 3.6 \end{pmatrix}$$

yields the least squares line $\mathcal{L}$ as

$$\widehat{y}(x) = \widehat{\alpha} + \widehat{\beta}x = 19.5 + 3.6x,$$

which in turn provides a sales estimate for any year x—e.g., $\widehat{y}(5) = \$375,000$ is the estimated sales for year five. To get a feel for how well the line $\mathcal{L}$ explains the observed sales over time, set $\widehat{\mathbf{y}} = \mathbf{A}\widehat{\mathbf{x}}$, and compute the coefficient of determination from (4.4.8) to be

$$r^2 = \frac{\|\widehat{\mathbf{y}} - \mu\mathbf{e}\|_2^2}{\|\mathbf{y} - \mu\mathbf{e}\|_2^2} = \frac{64.8}{65} \approx .996923.$$

Thus about 99.7% of the variation in sales over time is explained by $\mathcal{L}$, or equivalently, only about .3% of the variation in sales over time is not explained by $\mathcal{L}$. This suggests that the least squares model can be a good predictor of future sales, assuming of course that the trend continues to hold.

Vector Space Theory of Least Squares

Viewing concepts from more than one perspective generally produces a deeper understanding, and this is particularly true for the theory of least squares. While the classical calculus-based theory of least squares as discussed earlier can be extended to cover more general situations, it is generally replaced by a more intuitive development based on vector space geometry. This approach not only produces a cleaner theory, but it also brings the entire least squares picture into sharper focus. Rather than fitting a straight line to a data set of ordered pairs, more general least squares concerns the following problem.

- Given $\mathbf{A} \in \mathbb{F}^{m\times n}$ and $\mathbf{y} \in \mathbb{F}^m$, find a vector $\widehat{\mathbf{x}} \in \mathbb{F}^n$ such that $\mathbf{A}\widehat{\mathbf{x}}$ is as close to $\mathbf{y}$ as possible in the sense that $\|\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}\|_2^2 = \min_{\mathbf{x}\in\mathbb{F}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$, or equivalently, $\|\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}\|_2 = \min_{\mathbf{x}\in\mathbb{F}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2$.

Since $\mathbf{A}\widehat{\mathbf{x}}$ is always a vector in $R(\mathbf{A})$, the problem boils down to finding the vector $\mathbf{p} \in R(\mathbf{A})$ that is closest to $\mathbf{y}$. The closest point theorem on page 463 solves this problem because it guarantees that $\mathbf{p} = \mathbf{P}_{R(\mathbf{A})}\mathbf{y}$, where $\mathbf{P}_{R(\mathbf{A})}$ is the orthogonal projector onto $R(\mathbf{A})$. Figure 4.4.3 below illustrates the situation in $\mathbb{R}^3$.

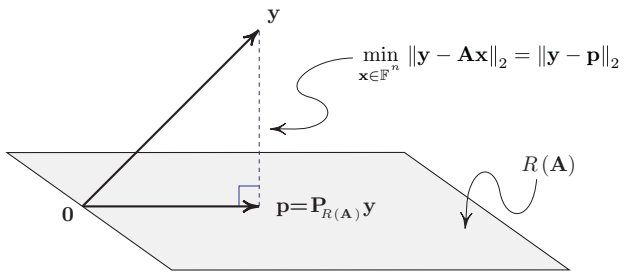


FIGURE 4.4.3: PROJECTION ONTO $R(\mathbf{A})$

Therefore, $\widehat{\mathbf{x}}$ is a vector such that $\|\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}\|_2^2$ is minimal if and only if

$$\mathbf{A}\widehat{\mathbf{x}} = \mathbf{p} = \mathbf{P}_{R(\mathbf{A})}\mathbf{y}.$$

However, this is just the system of normal equations $\mathbf{A}^*\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^*\mathbf{y}$ in disguised form[†] because

$$\begin{aligned}\mathbf{A}\widehat{\mathbf{x}} = \mathbf{P}_{R(\mathbf{A})}\mathbf{y} &\Longleftrightarrow \mathbf{P}_{R(\mathbf{A})}\mathbf{A}\widehat{\mathbf{x}} = \mathbf{P}_{R(\mathbf{A})}\mathbf{y} \Longleftrightarrow \mathbf{P}_{R(\mathbf{A})}(\mathbf{A}\widehat{\mathbf{x}} - \mathbf{y}) = \mathbf{0} \\ &\Longleftrightarrow (\mathbf{A}\widehat{\mathbf{x}} - \mathbf{y}) \in N\left(\mathbf{P}_{R(\mathbf{A})}\right) = R(\mathbf{A})^{\perp} = N(\mathbf{A}^*) \\ &\Longleftrightarrow \mathbf{A}^*(\mathbf{A}\widehat{\mathbf{x}} - \mathbf{y}) = \mathbf{0} \Longleftrightarrow \mathbf{A}^*\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^*\mathbf{y}.\end{aligned}$$

In summary, this means that the definition of a general least squares solution can be stated in any one of three equivalent ways.

4.4.3. Definition. A ***least squares solution*** for a system of linear equations $\mathbf{A}_{m\times n}\mathbf{x} = \mathbf{y}$ (possibly inconsistent) is defined to be a vector $\widehat{\mathbf{x}} \in \mathbb{F}^n$ that satisfies any one of the following three equivalent statements in which $\mathbf{P}_{R(\mathbf{A})}$ is the orthogonal projector onto $R(\mathbf{A})$.

- $\|\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}\|_2^2 = \min_{\mathbf{x}\in\mathbb{F}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ (4.4.9)
- $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{p} = \mathbf{P}_{R(\mathbf{A})}\mathbf{y}$ (the projection equation) (4.4.10)
- $\mathbf{A}^*\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^*\mathbf{y}$ (the normal equations) (4.4.11)

Note that the 2×2 system of normal equations in (4.4.4) on page 479 is just a special case of the more general system of normal equations (4.4.11) that results from the vector space theory.

Caution! The statements in (4.4.9)–(4.4.11) are the theoretical foundations for least squares theory, but they are generally not used for practical floating-point computation. Explicitly forming the product $\mathbf{A}^*\mathbf{A}$ and then solving the normal equations is ill-advised because if κ is the two-norm condition number for $\mathbf{A}$, then κ^2 the two-norm condition number for $\mathbf{A}^*\mathbf{A}$, (see Exercise 3.5.21 on page 377), so any sensitivities to small perturbations (e.g., rounding error) that are present in the underlying problem are magnified by computing $\mathbf{A}^*\mathbf{A}$ (see in Exercise 2.8.8 on page 242). Stable algorithms generally involve orthogonal reduction techniques that are discussed later in the text

[†] Note that this discussion allows for complex matrices whereas earlier discussions were restricted to real matrices. This is because traditional linear least squares analysis is almost always in the context of real numbers, but more general least squares applications can involve complex matrices.

All Least Squares Solutions

If $rank\,(\mathbf{A}_{m\times n}) = n$, then $(\mathbf{A}^*\mathbf{A})_{n\times n}$ is nonsingular, so the system of normal equations (4.4.11) yields a unique least squares solution given by

$$\widehat{\mathbf{x}} = (\mathbf{A}^*\mathbf{A})^{-1}\,\mathbf{A}^*\mathbf{y}.$$

But unlike the traditional problem on page 478, $\mathbf{A}$ need not have full column rank, in which case there are infinitely many least squares solutions. The set of all least squares solutions is the complete solution set for the projection equation $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{p}$ in (4.4.10), and Theorem 2.5.7 on page 208 ensures that this set is

$$\mathcal{S} = \mathbf{x}_{\text{part}} + N\,(\mathbf{A}),$$

where $\mathbf{x}_{\text{part}}$ is a particular solution of $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{p}$. A convenient particular solution is the pseudo inverse solution $\mathbf{x}_{\text{part}} = \mathbf{A}^\dagger\mathbf{y}$ because

$$\mathbf{A}(\mathbf{A}^\dagger\mathbf{y}) = (\mathbf{A}\mathbf{A}^\dagger)\mathbf{y} = \mathbf{P}_{R(\mathbf{A})}\mathbf{y} = \mathbf{p} \quad \text{(recall (4.3.17) on page 461).}$$

Therefore, the set of all least squares solutions for a general system $\mathbf{A}\mathbf{x} = \mathbf{y}$ is

$$\mathcal{S} = \mathbf{A}^\dagger\mathbf{y} + N\,(\mathbf{A}). \tag{4.4.12}$$

Note that $\mathcal{S}$ also the solution set for $\mathbf{A}\mathbf{x} = \mathbf{y}$ when this system is consistent because if $\mathbf{y} \in R\,(\mathbf{A})$, then $\mathbf{A}\mathbf{A}^\dagger\mathbf{y} = \mathbf{P}_{R(\mathbf{A})}\mathbf{y} = \mathbf{y}$.

Not only is $\mathbf{A}^\dagger\mathbf{y}$ a particular least squares solution, it is the unique minimal 2-norm solution among all least squares solutions. This follows from (4.4.12) because if $\mathbf{z}$ is any other least squares solution, then $\mathbf{z} = \mathbf{A}^\dagger\mathbf{y} + \mathbf{h}$, where $\mathbf{h} \in N\,(\mathbf{A}) = R\,(\mathbf{A}^*)^\perp = R\left(\mathbf{A}^\dagger\right)^\perp$ (recall (4.3.17), page 461), and hence the Pythagorean theorem (page 40) yields

$$\|\mathbf{z}\|_2^2 = \left\|\mathbf{A}^\dagger\mathbf{y} + \mathbf{h}\right\|_2^2 = \left\|\mathbf{A}^\dagger\mathbf{y}\right\|_2^2 + \|\mathbf{h}\|_2^2 \geq \left\|\mathbf{A}^\dagger\mathbf{y}\right\|_2^2,$$

with equality holding if and only if $\mathbf{h} = \mathbf{0}$—i.e., if and only if $\mathbf{z} = \mathbf{A}^\dagger\mathbf{y}$. These observations are summarized in the following theorem.

4.4.4. Theorem. Let $\mathbf{A}_{m\times n}\mathbf{x} = \mathbf{y}$ be a general system of linear equations.

- The set of all least squares solutions is $\mathcal{S} = \mathbf{A}^\dagger\mathbf{y} + N\,(\mathbf{A})$.
  - $\widehat{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{y}$ is the minimal 2-norm least squares solution.
- If the system is consistent, then its solution set is $\mathcal{S} = \mathbf{A}^\dagger\mathbf{y} + N\,(\mathbf{A})$.
  - $\widehat{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{y}$ is the minimal 2-norm solution.
- In either case, there is a unique solution (or least squares solution) if and only if $rank\,(\mathbf{A}) = n$, and it is given by

$$\widehat{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{y} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{y}.$$

Linear Regression

The traditional least squares problem of fitting data points (x_i, y_i) to a straight line as discussed on page 478 becomes the statistical theory of *linear regression* (also called *multiple regression*) when the goal is to relate a random variable y that cannot be observed exactly to a linear combination of two or more mathematical variables $x_1, x_2, \ldots, x_n$ that are not subject to error or variation and can be exactly measured or observed (e.g., $x_1 =$ the precise month of the year, $x_2 =$ the exact time of day, $x_3 =$ your current age, etc) together with another random variable ϵ such that

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon,$$

in which the parameters $\beta_0, \beta_1, \ldots, \beta_n$ are unknown constants. The role of ϵ is to account for the fact that y cannot be observed or measured exactly, or that other factors (e.g., simplifying assumptions or modeling errors) are not considered, but the effects of all of these errors[†] "average out" to zero in the sense that $E[\epsilon] = 0$, where $E[\star]$ denotes expected value (or mean). In other words, the regression assumption is that the mean value of y at each point where $(x_1, x_2, \ldots, x_n)$ can be observed is given by

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n. \tag{4.4.13}$$

Estimating the unknown parameters β_i involves making a series of m measurements or observations of y and $(x_1, x_2, \ldots, x_n)$ and hypothesizing that

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_n x_{nj} + \epsilon_j, \quad i = 1, 2, \ldots, m, \tag{4.4.14}$$

where y_j and x_{ji} are the respective j^{th} observations of y and x_i, and where ϵ_j is a random error for which it assumed that $E[\epsilon_i] = 0$. This results in vectors and matrices

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}$$

such that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, or equivalently $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\epsilon}$. An estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is provided by the general theory of least squares by taking $\widehat{\boldsymbol{\beta}}$ to be a vector such that $\left\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\|_2^2$ is minimal, or equivalently, $\widehat{\boldsymbol{\beta}}$ is a solution to the system of

[†] The difference between a measurement (or observation) error and a modeling error may be significant to an experimentalist, but mathematicians and statisticians generally do not make a distinction because they are equivalent from a mathematical standpoint.

normal equations $\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$. Consequently, the estimate of the mean value of y for any given set of values for $(x_1, x_2, \ldots, x_n)$ is

$$\widehat{E[y]} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_n x_n, \tag{4.4.15}$$

where $\widehat{\beta}_i$ is the least squares estimate for β_i. For the least squares estimate $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, exactly the same analysis that led to the coefficient of determination on page 481 holds for the more general case of multiple regression, so

$$r^2 = \frac{\|\widehat{\mathbf{y}} - \mu\mathbf{e}\|_2^2}{\|\mathbf{y} - \mu\mathbf{e}\|_2^2}, \quad \text{where} \quad \mu = \mu_{\mathbf{y}}, \tag{4.4.16}$$

can again be used to gauge the proportion of the observed vector $\mathbf{y}$ that is explained by the regression model, and thus it measures the goodness of fit.

The Gauss–Markov theorem on page 491 says that under reasonable assumptions about the random error $\boldsymbol{\epsilon}$, the least squares estimates $\widehat{\boldsymbol{\beta}}$ for the $\boldsymbol{\beta}$ and the estimate (4.4.15) for (4.4.13) is optimal in a particular sense. But before getting into this, it may be helpful to look at a simple example.

Example (Stale Pop)

Everyone knows that when the unused half of an open can of soda (or "pop" as it is called in Greeley Colorado) is put back into the refrigerator, it increasingly loses its palatability the longer it is left. For a particular brand (say Coca Cola) there can be several factors that influence this—e.g, the number of days an opened can remains in the refrigerator, the refrigerator's temperature, the original level of carbonation, amounts of ingredients such as high fructose corn syrup, artificial sweeteners, phosphoric acid, flavorings, etc. It is reasonable to conjecture that storage time and temperature are the primary factors, and other factors "average out," so to predict the loss of palatability, make a linear hypothesis of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \quad \text{where} \quad E[\epsilon] = 0$$

in which

$y =$ the percent of palatability lost compared to that of a fresh can as subjectively judged by a panel of tasters,

$x_1 =$ the number of days stored after opening,

$x_2 =$ the temperature ($^\circ C$) of the refrigerator during storage time.

In other words, the hypothesis is $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, and producing an estimate $\widehat{E[y]} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$ is accomplished by conducting experiments to determine least squares estimates for each $\widehat{\beta}_i$. The following table records the storage times and temperatures for nine experiments along with judgements of loss of palatability for each of these values.

x_1 = Storage Time (days)	1	1	1	2	2	2	3	3	3
x_2 = Storage Temp (^oC)	1	2	3	1	2	3	1	2	3
y = Palatability Loss (%)	15	16	20	17	19	22	20	23	25

The model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, were

$$\mathbf{y} = \begin{pmatrix} 15 \\ 16 \\ 20 \\ 17 \\ 19 \\ 22 \\ 20 \\ 23 \\ 25 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 2 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \\ 1 & 3 & 1 \\ 1 & 3 & 2 \\ 1 & 3 & 3 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_2 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \end{pmatrix}$$

with the assumption that $E[\epsilon_i] = 0$ for each i. Least squares estimates $\widehat{\beta}_i$ for the β_i's are obtained by solving the normal equations

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} \implies \begin{pmatrix} 9 & 16 & 18 \\ 18 & 42 & 36 \\ 18 & 36 & 42 \end{pmatrix}\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 177 \\ 371 \\ 369 \end{pmatrix} \implies \widehat{\boldsymbol{\beta}} = \begin{pmatrix} 9 \\ 17/6 \\ 5/2 \end{pmatrix}.$$

In this case the coefficient of determination (to three significant places) is

$$r^2 = \frac{\|\widehat{\mathbf{y}} - \mu\mathbf{e}\|_2^2}{\|\mathbf{y} - \mu\mathbf{e}\|_2^2} = 97.3,$$

so more than 97% of the variation $\mathbf{y}$ is explained by the regression model while less than 3% of the variation is not, and thus the least squares fit is pretty good. For example, the regression model predicts that a half can of opened Coke stored for three days in a refrigerator set at $4\,^oC$ is expected to lose about

$$\widehat{\beta}_0 + \widehat{\beta}_1(3) + \widehat{\beta}_2(4) = 9 + (7/16)(3) + (5/2)(4) = 27.5\%$$

of the palatability of an unopened can.

Least Squares Estimates are Optimal

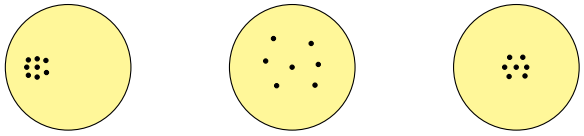
Drawing inferences about natural phenomena based upon physical observations and estimating characteristics of large populations by examining small samples are fundamental concerns of applied science. Numerical characteristics of a phenomenon or population are often called *parameters,* and the goal is to design functions or rules called *estimators* that use observations or samples to estimate parameters of interest. For example, the mean height h of all people is a parameter of the world's population, and one way of estimating h is to observe

the mean height of a sample of k people. In other words, if h_i is the height of the i^{th} person in a sample, then the function $\widehat{h}$ defined by

$$\widehat{h}(h_1, h_2, \ldots, h_k) = \frac{1}{k}\left(\sum_{i=1}^{k} h_i\right)$$

is an estimator for h. Moreover, $\widehat{h}$ is a *linear estimator* because $\widehat{h}$ is a linear function of the observations h_i.

Good estimators should possess at least two properties—they should be *unbiased* and they should have *minimal variance.* For example, consider estimating the center of a circle drawn on a wall by asking Larry, Moe, and Curly to each throw one dart at the circle. To decide which estimator is best, knowledge about each thrower's style is required. While being able to throw a tight pattern, it is known that Larry tends to have a left-hand bias in his style. Moe doesn't suffer from a bias, but he tends to throw a rather large pattern. However, Curly can throw a tight pattern without a bias. Typical patterns are shown below.



Larry Moe Curly

Although Larry has a small variance, he is a poor estimator because he is biased in the sense that his average is significantly different than the center. Moe and Curly are each unbiased estimators because they have an average that is the center, but Curly is clearly the preferred estimator because his variance is smaller than Moe's. In other words, Curly is the unbiased estimator of minimal variance.

4.4.5. Definition. An estimator $\widehat{\theta}$ (considered as a random variable) for a parameter θ is said to be ***unbiased*** when $E[\widehat{\theta}] = \theta$, and $\widehat{\theta}$ is called a ***minimum variance unbiased estimator*** for θ whenever $\mathrm{Var}[\widehat{\theta}] \leq \mathrm{Var}[\widehat{\phi}]$ for all unbiased estimators $\widehat{\phi}$ of θ.

These ideas make it possible to precisely articulate the sense in which least squares is optimal. Consider a linear hypothesis

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon,$$

in which y is a random variable that cannot be exactly observed (perhaps due to measurement error), the x_i's are mathematical variables whose values are

not subject to error or variation (they can be exactly observed or measured), and where ϵ is a random variable accounting for the error. As explained in the previous section, least squares estimates $\widehat{\beta}_i$ for β_i are obtained by observing values y_j of y at m different points $(x_{j1}, x_{j2}, \ldots, x_{jn}) \in \mathbb{R}^n$, where x_{ji} is the value of x_i to be used when making the j^{th} observation. This produces the *linear model*

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_n x_{nj} + \epsilon_j, \quad i = 1, 2, \ldots, m, \tag{4.4.17}$$

in which ϵ_j is a random variable accounting for the j^{th} observation or measurement error. Thus y_j is also a random variable. A standard assumption is that observation errors are not correlated with each other, but they have a common variance σ^2 (not necessarily known) and a zero mean. In other words, it is assumed that [†]

$$E[\epsilon_i] = 0 \text{ for each } i \quad \text{and} \quad \mathrm{Cov}[\epsilon_i, \epsilon_j] = \begin{cases} \sigma^2 & \text{when } i = j, \\ 0 & \text{when } i \neq j. \end{cases}$$

$$\text{If } \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix},$$

then the equations in (4.4.17) can be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. In practice, $m > n$ points $(x_{j1}, x_{j2}, \ldots, x_{jn}) \in \mathbb{R}^n$ at which observations y_j are made can almost always be selected to insure that $\mathbf{X}$ has full column rank (see Exercise 4.4.11 for the rank-deficient case), so the complete statement of a standard linear model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{such that} \begin{cases} rank\left(\mathbf{X}_{m\times(n+1)}\right) = n+1, \\ E[\boldsymbol{\epsilon}] = \mathbf{0} \quad (\text{so } E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}), \\ \mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I} = \mathrm{Cov}[\mathbf{y}] \quad (\text{so } \mathrm{Var}[\epsilon_i] = \sigma^2 = \mathrm{Var}[y_i]), \end{cases} \tag{4.4.18}$$

in which

$$E[\boldsymbol{\epsilon}] = \begin{pmatrix} E[\epsilon_1] \\ E[\epsilon_2] \\ \vdots \\ E[\epsilon_m] \end{pmatrix} \text{ and } \mathrm{Cov}[\boldsymbol{\epsilon}] = \begin{pmatrix} \mathrm{Cov}[\epsilon_1, \epsilon_1] & \mathrm{Cov}[\epsilon_1, \epsilon_2] & \cdots & \mathrm{Cov}[\epsilon_1, \epsilon_m] \\ \mathrm{Cov}[\epsilon_2, \epsilon_1] & \mathrm{Cov}[\epsilon_2, \epsilon_2] & \cdots & \mathrm{Cov}[\epsilon_2, \epsilon_m] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[\epsilon_m, \epsilon_1] & \mathrm{Cov}[\epsilon_m, \epsilon_2] & \cdots & \mathrm{Cov}[\epsilon_m, \epsilon_m] \end{pmatrix}.$$

A primary goal is to determine the best (i.e., minimum variance) linear (linear function of the y_i's) unbiased estimators for the components of $\boldsymbol{\beta}$. Gauss realized that this is precisely what the theory of least squares provides.

[†] Recall from elementary probability that for random variables A, B and a constants a, b,

- $E[aA + B] = aE[A] + E[B]$ (i.e., expectation is linear),
- $\mathrm{Var}[A] = E\left[(A - \mu_A)^2\right] = E[A^2] - \mu_A^2$,
- $\mathrm{Cov}[AB] = E[(A - \mu_A)(B - \mu_B)] = E[AB] - \mu_A\mu_B$,
- $\mathrm{Var}[aA + bB] = a^2\mathrm{Var}[A] + b^2\mathrm{Var}[B]$ when $\mathrm{Cov}[AB] = 0$.

4.4.6. Theorem. (Gauss–Markov† Theorem) For the linear model (4.4.18), the minimum variance linear unbiased estimator $\widehat{\beta}_i$ for β_i is the i^{th} component of $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^\dagger\mathbf{y}$. In other words, the best linear unbiased estimator for $\boldsymbol{\beta}$ is the least squares solution of $\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{y}$.

- Moreover, for a row $\mathbf{t}^T$ of constants, $\mathbf{t}^T\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator for linear combination $\mathbf{t}^T\boldsymbol{\beta}$. (4.4.19)

Proof. It is clear that $\widehat{\boldsymbol{\beta}} = \mathbf{X}^\dagger\mathbf{y}$ is a linear estimator because each component $\widehat{\beta}_i = \sum_k [\mathbf{X}^\dagger]_{ik}\, y_k$ is a linear function of the observations y_k. To see that $\widehat{\boldsymbol{\beta}}$ is unbiased, use $E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{X}^\dagger = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ (see page 191) to write

$$E\left[\widehat{\boldsymbol{\beta}}\right] = E\left[\mathbf{X}^\dagger\mathbf{y}\right] = \mathbf{X}^\dagger E[\mathbf{y}] = \mathbf{X}^\dagger\mathbf{X}\boldsymbol{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

To argue that $\widehat{\boldsymbol{\beta}} = \mathbf{X}^\dagger\mathbf{y}$ has minimal variance among all linear unbiased estimators for $\boldsymbol{\beta}$, let $\boldsymbol{\beta}^*$ be an arbitrary linear unbiased estimator for $\boldsymbol{\beta}$. Linearity of $\boldsymbol{\beta}^*$ implies the existence of a (constant) $(n+1)\times m$ matrix $\mathbf{L}$ such that $\boldsymbol{\beta}^* = \mathbf{L}\mathbf{y}$ so that

$$\mathrm{Var}[\beta_i^*] = \mathrm{Var}[\mathbf{L}_{i*}\mathbf{y}] = \mathrm{Var}\left[\sum_{k=1}^{m} l_{ik}y_k\right] = \sigma^2\sum_{k=1}^{m} l_{ik}^2 = \sigma^2\left\|\mathbf{L}_{i*}\right\|_2^2$$

(see the variance formula in the previous footnote). Unbiasedness ensures that $\boldsymbol{\beta} = E[\boldsymbol{\beta}^*] = E[\mathbf{L}\mathbf{y}] = \mathbf{L}E[\mathbf{y}] = \mathbf{L}\mathbf{X}\boldsymbol{\beta}$ for all $\boldsymbol{\beta}\in\mathbb{R}^{n+1}$, and hence $\mathbf{L}\mathbf{X} = \mathbf{I}_{n+1}$ (Exercise 1.7.10, page 65). Therefore, $\mathrm{Var}[\beta_i^*]$ is minimal if and only if $\mathbf{L}_{i*}$ is the minimum norm solution for $\mathbf{z}^T$ in the left-handed system $\mathbf{z}^T\mathbf{X} = \mathbf{e}_i^T$. In general, the unique minimum norm solution is given by $\mathbf{z}^T = \mathbf{e}_i^T\mathbf{X}^\dagger = \mathbf{X}_{i*}^\dagger$ (Theorem 4.4.4, page 485), and thus $\mathrm{Var}[\beta_i^*]$ is minimal for each i if and only if $\mathbf{L}_{i*} = \mathbf{X}_{i*}^\dagger$, or equivalently, $\mathbf{L} = \mathbf{X}^\dagger$. Therefore, the (unique) minimal variance linear unbiased estimator for $\boldsymbol{\beta}$ is $\mathbf{X}^\dagger\mathbf{y} = \widehat{\boldsymbol{\beta}}$. The proof of (4.4.19) follows along the same lines, so details are left to the reader. ∎

† Gauss is generally credited with realizing this theorem in 1821, but historians are ambivalent about Markov's contribution. Andrie Andreevich (or Andrey Andreyevich) Markov (1856–1922) was a slow starter in school at Petrograd (now St Petersburg), Russia, but he eventually found his stride in studying mathematics and became a student of Pafnuty Chebyshev who stimulated his interest in probability. This led to Markov's development of "Markov chains," which subsequently launched the theory of stochastic processes. Markov preferred the rigorous side of probability theory, and he is said originally to have had a negative attitude toward statistics—he judged it strictly from a mathematical point of view. But when his views later softened, his interests merged to produce what eventually became known as "mathematical statistics." However, in the historical anthology *Statisticians of the Centuries, 2001st Edition,* Eugene Seneta suggests that while Markov had an interest in statistical linear models, it may be inappropriate for his name to be attached to this theorem.

Least Squares Curve Fitting

When a set of observations $\mathcal{D} = \{(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_m, y_m)\}$ does not follow a linear trend, attempting to fit the data to a straight line as described on page 478 is not productive, so a more general approach is to fit the data to a curve defined by a polynomial

$$p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_n x^n$$

with a specified degree $n < m$ that comes as close as possible in the sense of ordinary least squares.

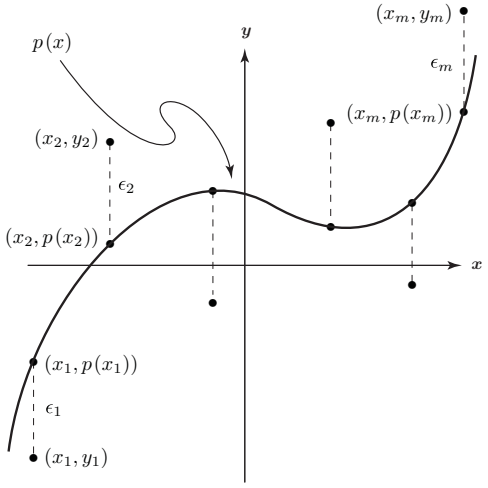


FIGURE 4.4.4: LEAST SQUARES POLYNOMIAL

The same assumptions described for ordinary least squares problems as discussed on page 478 remain in effect, and the analysis is identical to that described on page 483. For the ϵ_i's indicated in Figure 4.4.4, the objective is to minimize the sum of squares

$$\sum_{i=1}^{m} \epsilon_i^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = \sum_{i=1}^{m} \left(y_i - p(x_i)\right)^2 = (\mathbf{y} - \mathbf{A}\mathbf{x})^T (\mathbf{y} - \mathbf{A}\mathbf{x}), \tag{4.4.20}$$

where

$$\mathbf{A} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{pmatrix},\ \mathbf{x} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix},\ \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}\ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix} = \mathbf{y} - \mathbf{A}\mathbf{x}.$$

In other words, the least squares polynomial of degree n is obtained from the least squares solution associated with the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ because it was demonstrated on page 484 that a vector $\widehat{\mathbf{x}}$ such that $\|\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}\|_2^2$ is minimal must

satisfy the normal equations $\mathbf{A}^T\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^T\mathbf{y}$. The least squares polynomial $p(x) = \widehat{\alpha}_0 + \widehat{\alpha}_1 x + \cdots + \widehat{\alpha}_n x^n$ defined by $\widehat{\mathbf{x}} = \begin{pmatrix} \widehat{\alpha}_0 \\ \widehat{\alpha}_1 \\ \vdots \\ \widehat{\alpha}_n \end{pmatrix}$ is unique provided that $x_i \neq x_j$ for all $i \neq j$ because $\mathbf{A}$ is a Vandermonde matrix, and it is shown on page 60 that all such matrices have full column rank, which in turn makes $\mathbf{A}^T\mathbf{A}$ nonsingular so that

$$\widehat{\mathbf{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$$

is the unique least squares solution.

Example (Missile Tracking)

A missile is fired from enemy territory, and its position in flight is observed by radar tracking devices at the following positions.

x =Position down range (miles)	0	250	500	750	1000
y =Height (miles)	0	8	15	19	20

Suppose that intelligence sources indicate that enemy missiles are programmed to follow a parabolic flight path—a fact that seems to be consistent with the trend suggested by plotting the observations as shown below.

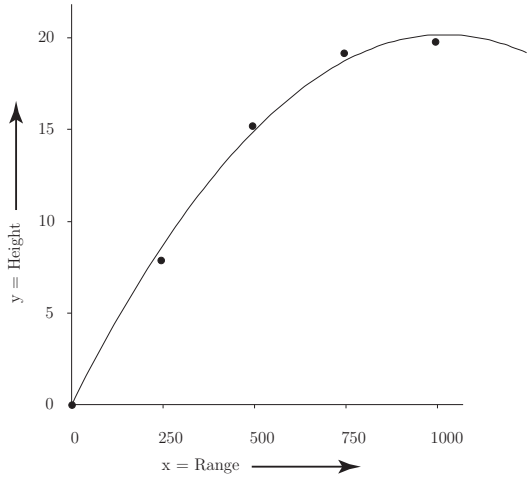


FIGURE 4.4.5: MISSILE OBSERVATIONS

Problem: Where is the missile expected to land?

Solution: Determine the parabola $p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ that best fits the observed data in the ordinary least squares sense, and then estimate where the missile will land by finding the roots of p to determine where the parabola crosses the horizontal axis. In its raw form the problem will involve numbers

having relatively large magnitudes in conjunction with relatively small ones, so it is better to first scale the data by considering one unit to be 1000 miles. For

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & .25 & .0625 \\ 1 & .5 & .25 \\ 1 & .75 & .5625 \\ 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} 0 \\ .008 \\ .015 \\ .019 \\ .02 \end{pmatrix},$$

the aim, as described on page 492, is to find a value $\widehat{\mathbf{x}}$ such that

$$\|\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}\|_2^2 = \min_{\mathbf{x}\in\mathbb{R}^3} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2,$$

which is equivalent to solving the normal equations

$$\mathbf{A}^T\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^T\mathbf{y} \implies \begin{pmatrix} 5 & 2.5 & 1.875 \\ 2.5 & 1.875 & 1.5625 \\ 1.875 & 1.5625 & 1.3828125 \end{pmatrix}\begin{pmatrix} \widehat{\alpha}_0 \\ \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} .062 \\ .04375 \\ .0349375 \end{pmatrix}$$

$$\implies \widehat{\mathbf{x}} = \begin{pmatrix} \widehat{\alpha}_0 \\ \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} -2.285714 \times 10^{-4} \\ 3.982857 \times 10^{-2} \\ -1.942857 \times 10^{-2} \end{pmatrix} \quad \text{(to seven places)}.$$

Thus the least squares parabola is

$$p(x) = -.0002285714 + .03982857x - .01942857x^2,$$

and the quadratic formula yields $x = 0.005755037$ and $x = 2.044245$ (to seven places). These are where $p(x)$ crosses the x-axis, so the estimated point of impact (after rescaling) is 2044.245 miles down range.† To get a sense of how good the fit is, compute

$$\widehat{\mathbf{y}} = \mathbf{A}\widehat{\mathbf{x}} = \begin{pmatrix} -2.285714 \times 10^{-4} \\ 8.514286 \times 10^{-3} \\ 1.482857 \times 10^{-2} \\ 1.871426 \times 10^{-2} \\ 2.017143 \times 10^{-2} \end{pmatrix} \quad \mu_{\widehat{\mathbf{y}}} = 0.0124 = \mu_{\mathbf{y}}$$

to evaluate the coefficient of determination from (4.4.16) on page 487 to be

$$r^2 = \frac{\|\widehat{\mathbf{y}} - \mu\mathbf{e}\|_2^2}{\|\mathbf{y} - \mu\mathbf{e}\|_2^2} = \frac{2.807428 \times 10^{-4}}{2.812000 \times 10^{-4}} = 0.9983743.$$

Therefore, about 99.84% the variation in $\mathbf{y}$ is explained by the least squares parabola, so the fit is pretty good, and people who are somewhere around 2044 miles down range should seek immediate shelter.

† Remember that the observations are not expected to lie exactly on the least squares curve, so $x = 0.005755037$ is just the least squares estimate of the launch point (the origin).

Least Squares vs. Lagrange Interpolation

For a given set of m points $\mathcal{D} = \{(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_m, y_m)\}$ in which the x_i's are distinct, it is established on page 160 that the Lagrange interpolation polynomial

$$\ell(x) = \sum_{i=1}^{m} \left(y_i \frac{\prod_{j \neq i}^{m} (x - x_j)}{\prod_{j \neq i}^{m} (x_i - x_j)} \right)$$

exactly passes through each point in $\mathcal{D}$. So why would one want to settle for a least squares fit when an exact fit is possible?

One answer is that in practical work the observations y_i are rarely exact due to small errors arising from imprecise measurements or from simplifying assumption, so the goal is to fit the *trend* of the observations and not the uncertain observations themselves. Furthermore, to exactly hit all data points, the interpolation polynomial $\ell(x)$ is usually forced to oscillate between or beyond the data points, and as m becomes larger the oscillations can become more pronounced. Consequently, $\ell(x)$ is generally not useful in making predictions beyond the observations. The missile tracking example on page 493 drives this point home.

The fourth-degree Lagrange interpolation polynomial for the five observations listed on page 493 is

$$\ell(x) = \frac{11}{375}x + \frac{17}{750000}x^2 - \frac{1}{18750000}x^3 + \frac{1}{46875000000}x^4.$$

It can be verified that $\ell(x_i) = y_i$ for each observation. As the graph in Figure 4.4.6 indicates, $\ell(x)$ has only one real nonnegative root, so it is worthless for predicting where the missile will land. This is characteristic of Lagrange interpolation.

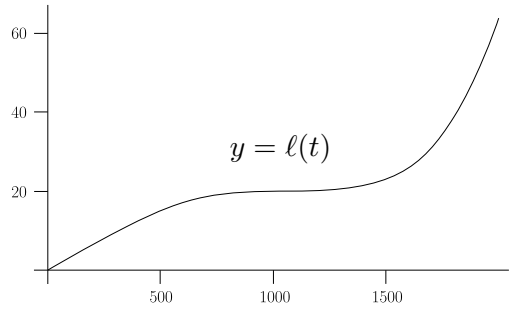


Figure 4.4.6: Interpolation polynomial for missile tracking data

Epilogue

It was mentioned on page 479 that Sir Francis Galton introduced the concept of linear regression, but he did not invented the theory of least squares. That honor belongs to Carl Gauss. While viewing a region in the Taurus constellation on January 1, 1801, Giuseppe Piazzi, an astronomer and director of the Palermo observatory, observed a small "star" that he had never seen before. As Piazzi and others continued to watch this new "star" (which was really an asteroid) they noticed that it was in fact moving, and they concluded that a new "planet" had been discovered—a really big deal back then. However, their new "planet" completely disappeared in the autumn of 1801. Well-known astronomers of the time joined the search to relocate the lost "planet," but all efforts were in vain.

In September of 1801 Gauss decided to take up the challenge of finding this lost "planet." Gauss allowed for the possibility of an elliptical orbit rather than constraining it to be circular—which was an assumption of the others—and he proceeded to develop the method of least squares. By December the task was completed, and Gauss informed the scientific community not only where the lost "planet" was located, but he also predicted its position at future times. They looked, and it was exactly where Gauss had predicted it would be! The asteroid was named *Ceres*, and Gauss's contribution was recognized by naming another minor asteroid *Gaussia.*

This extraordinary feat of locating a tiny and distant heavenly body from apparently insufficient data astounded the scientific community. Furthermore, Gauss refused to reveal his methods, and there were those who even accused him of sorcery. These events led directly to Gauss's fame throughout the entire European community, and they helped to establish his reputation as a mathematical and scientific genius of the highest order.

Gauss waited until 1809, when he published his *Theoria Motus Corporum Coelestium In Sectionibus Conicis Solem Ambientium,* to systematically develop the theory of least squares and his methods of orbit calculation. This was in keeping with Gauss's philosophy to publish nothing but well-polished work of lasting significance. When criticized for not revealing more motivational aspects in his writings, Gauss remarked that architects of great cathedrals do not obscure the beauty of their work by leaving the scaffolds in place after the construction has been completed. Gauss's theory of least squares has indeed proven to be a great mathematical cathedral of lasting beauty and significance.

Exercises for section 4.4

4.4.1. Consider the ordinary least squares problem of fitting a straight line $y = \alpha + \beta x$ to m non-colinear data points (x_i, y_i). Let $\widehat{\mathbf{x}}$ be the least

squares solution such that $\mathbf{A}^T\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^T\mathbf{y}$ in which $\mathbf{A} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix}$, $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \notin R(\mathbf{A})$, and $rank(\mathbf{A}) = 2$.

(a) Show that $\widehat{\mathbf{x}} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \dfrac{1}{\Delta}\begin{pmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ m\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right) \end{pmatrix}$, where $\Delta = m\sum x_i^2 - \left(\sum x_i\right)^2 = m\,\|\mathbf{x} - \mu_{\mathbf{x}}\mathbf{e}\|_2^2$ in which $\mathbf{e}$ is the vector of 1's and $\mu_\star$ denotes the mean.

(b) Prove that $(\mu_{\mathbf{x}}, \mu_{\mathbf{y}})$ always lies on the least squares line.

Hint: Show $\beta = \dfrac{(\mathbf{x} - \mu_{\mathbf{x}}\mathbf{e})^T(\mathbf{y} - \mu_{\mathbf{y}}\mathbf{e})}{\|\mathbf{x} - \mu_{\mathbf{x}}\mathbf{e}\|_2^2}$ and $\alpha = \mu_{\mathbf{y}} - \beta\mu_{\mathbf{x}}$.

(c) Show that $\beta = \dfrac{s_{\mathbf{xy}}}{s_{\mathbf{x}}^2} = \dfrac{\mathrm{Cov}[\mathbf{x},\mathbf{y}]}{\mathrm{Var}[\mathbf{x}]} = r_{\mathbf{xy}}\dfrac{s_{\mathbf{x}}}{s_{\mathbf{x}}^2}$, where $s_{\mathbf{x}}$, $s_{\mathbf{xy}}$, and $r_{\mathbf{xy}}$ are the respective sample standard deviation, covariance, and correlation as defined in (1.6.4), (1.6.10), and (1.6.11) on pages 44–46.

4.4.2. For the least squares problem described in Exercise 4.4.1, prove that $(\mathbf{y} - \widehat{\mathbf{y}}) \perp (\widehat{\mathbf{y}} - \mu_{\mathbf{y}}\mathbf{e})$, where $\widehat{\mathbf{y}} = \mathbf{A}\widehat{\mathbf{x}}$ in which $\widehat{\mathbf{x}}$ is the associated squares solution, and $\mu = \mu_{\mathbf{y}} = \mu_{\widehat{\mathbf{y}}}$.

4.4.3. For the ordinary least squares problem, let ϵ_i be the i^{th} error (the vertical deviation from the least squares line) as depicted in Figure 4.4.1 on page 478. Prove that $\sum_i \epsilon_i = 0$.

4.4.4. Hooke's law says that the displacement y of an ideal spring is proportional to the force x that is applied—i.e., $y = kx$ for some constant k. Consider a spring in which k is unknown. Various masses are attached, and the resulting displacements shown in Figure 4.4.7 are observed. Using these observations, determine the least squares estimate for k.

x (lb)	y (in)
5	11.1
7	15.4
8	17.5
10	22.0
12	26.3

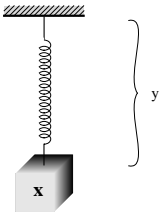


FIGURE 4.4.7: HANGING SPRING

4.4.5. Show that the slope of the line that passes through the origin in $\mathbb{R}^2$ and comes closest in the least squares sense to passing through the points $\{(x_1, y_1),\ (x_2, y_2), \ldots, (x_m, y_m)\}$ is given by $\beta = \sum_i x_i y_i / \sum_i x_i^2$.
Caution! The result of Exercise 4.4.1 does not apply to this case.

4.4.6. A small company has been in business for three years and has recorded annual profits (in thousands of dollars) as follows.

Year	1	2	3
Sales	7	4	3

Assuming that there is a linear trend in the declining profits, predict the year and the month in which the company begins to lose money.

4.4.7. An economist hypothesizes that the change (in dollars) in the price of a loaf of bread is primarily a linear combination of the change in the price of a bushel of wheat and the change in the minimum wage. That is, if B is the change in bread prices, W is the change in wheat prices, and M is the change in the minimum wage, then $B = \alpha W + \beta M$. Suppose that for three consecutive years the change in bread prices, wheat prices, and the minimum wage are as shown below.

	Year 1	Year 2	Year 3
B	+\$1	+\$1	+\$1
W	+\$1	+\$2	0\$
M	+\$1	0\$	−\$1

Use the theory of least squares to estimate the change in the price of bread in Year 4 if wheat prices and the minimum wage each fall by \$1.

4.4.8. Consider the problem of predicting the amount of weight that a pint of ice cream loses when it is stored at low temperatures. There are many factors that may contribute to weight loss—e.g., storage temperature, storage time, humidity, atmospheric pressure, butterfat content, the amount of corn syrup, the amounts of guar gum, carob bean gum, locust bean gum, cellulose gum, and the lengthly list of other additives and preservatives sometimes used. Conjecture that storage time and temperature are the primary factors, so to predict weight loss make a linear hypothesis of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where $y =$ weight loss (grams), $x_1 =$ storage time (weeks), $x_2 =$ storage temperature (^oF), and ϵ is a random variable to account for all other factors. Assume that $E[\epsilon] = 0$, so the expected weight loss at each point (x_1, x_2) is $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. An experiment in which values for weight loss are measured for various values of storage time and temperature as shown below.

Time (weeks)	1	1	1	2	2	2	3	3	3
Temp (^oF)	−10	−5	0	−10	−5	0	−10	−5	0
Loss (grams)	.15	.18	.20	.17	.19	.22	.20	.23	.25

Using this data, estimate the expected weight loss of a pint of ice cream that is stored for nine weeks at a temperature of -35^oF. Then use the coefficient of determination to gauge the goodness of fit.

4.4.9. After studying a certain type of cancer, a researcher hypothesizes that in the short run the number (y) of malignant cells in a particular tissue grows exponentially with time (x). That is, $y = \beta_0 e^{\beta_1 t}$. Determine least squares estimates for the parameters β_0 and β_1 from the researcher's observed data given below.

t (days)	1	2	3	4	5
y (cells)	16	27	45	74	122

Hint: What common transformation converts an exponential function into a linear function?

4.4.10. For $\mathbf{A} \in \mathbb{F}^{m\times n}$ and $\mathbf{y} \in \mathbb{F}^m$, prove that $\mathbf{x}_2$ is a least squares solution for $\mathbf{Ax} = \mathbf{y}$ if and only if $\mathbf{x}_2$ is part of a solution to the larger system

$$\begin{pmatrix} \mathbf{I}_{m\times m} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}.$$

Note: It is not uncommon to encounter least squares problems in which $\mathbf{A}$ is large and sparse (mostly zero entries). For these situations the system above may contain significantly fewer nonzero entries than the system of normal equations thereby helping to mitigate memory requirements, and it circumvents the need to explicitly form the product $\mathbf{A}^*\mathbf{A}$ that has inherent numerical sensitivities as explained on page 484.

4.4.11. Rank Deficient Models. In multiple regression models $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ as described on page 486 it can happen that the matrix $\mathbf{X}_{m\times(n+1)}$ is rank deficient (i.e., $rank\,(\mathbf{X}) < n+1$). Consequently, the normal equations $\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$ do not have a unique solution so that at any given point $(x_1, x_2, \ldots, x_n)$, there are infinitely many different estimates

$$\widehat{E[y]} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_n x_n.$$

To remedy the situation, points at where estimates are made must be restricted. Prove that if $\mathbf{t} = \begin{pmatrix} 1 \\ t_1 \\ \vdots \\ t_n \end{pmatrix} \in R\left(\mathbf{X}^T\right)$, and if $\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_n \end{pmatrix}$ is any least squares solution, then the estimate defined by

$$\widehat{E[y]} = \widehat{\beta}_0 + \widehat{\beta}_1 t_1 + \cdots + \widehat{\beta}_n t_n = \mathbf{t}^T\widehat{\boldsymbol{\beta}}$$

is unique in the sense that $\widehat{E[y]}$ is independent of which least squares solution $\widehat{\boldsymbol{\beta}}$ is used.

4.4.12. Using least squares, first fit the following data

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	2	7	9	12	13	14	14	13	10	8	4

with a line $y = \beta_0 + \beta_1 x$ and then fit the data with a quadratic function $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Determine which of these two curves best fits the data by computing the error sum of squares and the coefficient of determination in each case. (See the note in Exercise 4.4.13.)

4.4.13. Let $\widehat{\mathbf{x}}$ be the unique least squares problem associated with an inconsistent system $\mathbf{A}\mathbf{x} = \mathbf{y}$ in which $rank\,(\mathbf{A}_{m\times n}) = n$.

(a) Explain why the error sum of squares given in (4.4.5) on page 480 can be expressed as

$$\mathrm{SSE}_A = \sum_i \epsilon_i^2 = \|y - \mathbf{A}\widehat{\mathbf{x}}\|_2^2 = \|(\mathbf{I} - \mathbf{P}_A)\mathbf{y}\|_2^2, = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{P}_A\mathbf{y}$$

where $\mathbf{P}_A$ is the orthogonal projector onto $R\,(\mathbf{A})$.

(b) Prove that if $\mathbf{A}$ is augmented with a column $\mathbf{c}$ to produce $\mathbf{B} = [\mathbf{A}\,|\,\mathbf{c}]$ with $rank\,(\mathbf{B}) = n+1$, and if SSE_B is the error sum of squares for the least squares problem associated with $\mathbf{B}\widetilde{\mathbf{x}} = \mathbf{y}$, then $\mathrm{SSE}_B \leq \mathrm{SSE}_A$, with equality holding if and

only if $\mathbf{c}^T\mathbf{y} = \mathbf{c}^T\mathbf{A}\widehat{\mathbf{x}} = \mathbf{c}^T\mathbf{P}_A\mathbf{y}$, where $\widehat{\mathbf{x}}$ is the least squares solution for $\mathbf{Ax} = \mathbf{y}$. **Hint:** Recall Theorem 2.3.9 on page 165.

Note: This means that except for a rather special case, addition of variables to a linear regression model will reduce the error sum of squares, and the coefficient of determination $r^2 = 1 - \text{SSE/SST}$ will increase because $\text{SST} = \|\mathbf{y} - \mu\mathbf{e}\|_2^2$ depends only on $\mathbf{y}$. This was illustrated in Exercise 4.4.12.

4.4.14. Prove that for the standard linear model in (4.4.18), an unbiased estimator for σ^2 is given by

$$\widehat{\sigma}^2 = \frac{\mathbf{y}^T\mathbf{Q}\mathbf{y}}{m-n-1}, \quad \text{where} \quad \mathbf{Q} = \mathbf{I} - \mathbf{X}\mathbf{X}^\dagger.$$

Hint: Recall that the trace of an idempotent matrix is its rank (see Exercise 4.2.21, page 449), and use the fact that for a matrix $\mathbf{Z} = [z_{ij}]$ of random variables, $E[\mathbf{Z}]$ is defined to be the matrix whose (i, j)-entry is $E[z_{ij}]$.