Chapter 12

Linear Regression

In statistics, regression is the problem of characterizing the relation between a certain quantity of interest y, called the **response** or the **dependent variable**, to several observed variables x_1, x_2, ..., x_p, known as **covariates**, **features** or **independent variables**. For example, the response could be price of a house and the covariates could correspond to the extension, the number of rooms, the year it was built, etc. A regression model would describe how house prices are affected by all of these factors.

More formally, the main assumption in regression models is that the predictor is generated according to a function h applied to the features and then perturbed by some unknown noise z, which is often additive,

$$y = h(\vec{x}) + z. \tag{12.1}$$

The aim is to learn h from n examples of responses and their corresponding features

$$\left(y^{(1)}, \vec{x}^{\,(1)}\right), \left(y^{(2)}, \vec{x}^{\,(2)}\right), \ldots, \left(y^{(n)}, \vec{x}^{\,(n)}\right). \tag{12.2}$$

In this chapter we focus on the case where h is a linear function.

12.1 Linear models

If the regression function h in a model of the form 12.1 is linear, then the response is modeled as a linear combination of the predictors:

$$y^{(i)} = \vec{x}^{\,(i)\,T} \vec{\beta}^{*} + z^{(i)}, \quad 1 \leq i \leq n, \tag{12.3}$$

where $z^{(i)}$ is an entry of the unknown noise vector. The function is parametrized by a vector of weights $\vec{\beta}^{*} \in \mathbb{R}^p$. All we need to fit the linear model to the data is to estimate these weights.

Expressing the linear system (12.3) in matrix form yields the following representation of the linear-regression model

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} \vec{x}_1^{\,(1)} & \vec{x}_2^{\,(1)} & \cdots & \vec{x}_p^{\,(1)} \\ \vec{x}_1^{\,(2)} & \vec{x}_2^{\,(2)} & \cdots & \vec{x}_p^{\,(2)} \\ \cdots & \cdots & \cdots & \cdots \\ \vec{x}_1^{\,(n)} & \vec{x}_2^{\,(n)} & \cdots & \vec{x}_p^{\,(n)} \end{bmatrix} \begin{bmatrix} \vec{\beta}_1^{*} \\ \vec{\beta}_2^{*} \\ \cdots \\ \vec{\beta}_p^{*} \end{bmatrix} + \begin{bmatrix} z^{(1)} \\ z^{(2)} \\ \cdots \\ z^{(n)} \end{bmatrix}. \tag{12.4}$$

Equivalently,

$$\vec{y} = \mathcal{X}\vec{\beta}^{*} + \vec{z}, \tag{12.5}$$

where $\mathcal{X}$ is a $n \times p$ matrix containing the features, $\vec{y}$ contains the response and $\vec{z} \in \mathbb{R}^{n}$ represents the noise.

Example 12.1.1 (Linear model for GDP)**.** We consider the problem of building a linear model to predict the gross domestic product (GDP) of a state in the US from its population and unemployment rate. We have available the following data:

$$\begin{array}{l c c c}
 & \text{GDP} & \text{Population} & \text{Unemployment} \\
 & \text{(USD millions)} & & \text{rate (\%)} \\
\text{North Dakota} & 52\ 089 & 757\ 952 & 2.4 \\
\text{Alabama} & 204\ 861 & 4\ 863\ 300 & 3.8 \\
\text{Mississippi} & 107\ 680 & 2\ 988\ 726 & 5.2 \\
\text{Arkansas} & 120\ 689 & 2\ 988\ 248 & 3.5 \\
\text{Kansas} & 153\ 258 & 2\ 907\ 289 & 3.8 \\
\text{Georgia} & 525\ 360 & 10\ 310\ 371 & 4.5 \\
\text{Iowa} & 178\ 766 & 3\ 134\ 693 & 3.2 \\
\text{West Virginia} & 73\ 374 & 1\ 831\ 102 & 5.1 \\
\text{Kentucky} & 197\ 043 & 4\ 436\ 974 & 5.2 \\
\text{Tennessee} & ??? & 6\ 651\ 194 & 3.0
\end{array}$$

In this example, the GDP is the response, and the population and the unemployment rate are the features. Our goal is to fit a linear model to the data so that we can predict the GDP of Tennessee, using a linear model. We begin by centering and normalizing the data. The averages of the response and of the features are

$$\text{av}\left(\vec{y}\right) = 179\ 236, \qquad \text{av}\left(X\right) = \begin{bmatrix} 3\ 802\ 073 & 4.1 \end{bmatrix}. \tag{12.6}$$

The empirical standard deviations are

$$\text{std}\left(\vec{y}\right) = 396\ 701, \qquad \text{std}\left(X\right) = \begin{bmatrix} 7\ 720\ 656 & 2.80 \end{bmatrix}. \tag{12.7}$$

We subtract the average and divide by the standard deviations so that both the response and the features are centered and on the same scale,

$$\vec{y} = \begin{bmatrix} -0.321 \\ 0.065 \\ -0.180 \\ -0.148 \\ -0.065 \\ 0.872 \\ -0.001 \\ -0.267 \\ 0.045 \end{bmatrix}, \qquad X = \begin{bmatrix} -0.394 & -0.600 \\ 0.137 & -0.099 \\ -0.105 & 0.401 \\ -0.105 & -0.207 \\ -0.116 & -0.099 \\ 0.843 & 0.151 \\ -0.086 & -0.314 \\ -0.255 & 0.366 \\ 0.082 & 0.401 \end{bmatrix}. \tag{12.8}$$

To obtain the estimate for the GDP of Tennessee we fit the model

$$\vec{y} \approx X\vec{\beta}, \tag{12.9}$$

rescale according to the standard deviations (12.7) and recenter using the averages (12.6). The final estimate is

$$\vec{y}^{\,\text{Ten}} = \text{av}\left(\vec{y}\right) + \text{std}\left(\vec{y}\right) \left\langle \vec{x}^{\,\text{Ten}}_{\text{norm}}, \vec{\beta} \right\rangle \tag{12.10}$$

where $\vec{x}^{\,\text{Ten}}_{\text{norm}}$ is centered using $\text{av}\left(X\right)$ and normalized using $\text{std}\left(X\right)$. △

12.2 Least-squares estimation

To calibrate the linear regression model, we need to estimate the weight vector so that it yields a good fit to the data. We can evaluate the fit for a specific choice of $\vec{\beta} \in \mathbb{R}^p$ using the sum of the squares of the error,

$$\sum_{i=1}^{n} \left(y^{(i)} - \vec{x}^{\,(i)\,T} \vec{\beta} \right)^2 = \left| \left| \vec{y} - \mathcal{X}\vec{\beta} \right| \right|_2^2 . \tag{12.11}$$

The least-squares estimate $\vec{\beta}_{\text{LS}}$ is the vector of weights that minimizes this cost function,

$$\vec{\beta}_{\text{LS}} := \arg\min_{\vec{\beta}} \quad \left| \left| \vec{y} - \mathcal{X}\vec{\beta} \right| \right|_2 . \tag{12.12}$$

The least-squares cost function is convenient from a computational view, since it is convex and can be minimized efficiently (in fact, as we will see in a moment it has a closed-form solution). In addition, it has intuitive geometric and probabilistic interpretations. Figure 12.1 shows the linear model learnt using least squares in a simple example where there is just one feature ($p = 1$) and 40 examples ($n = 40$).

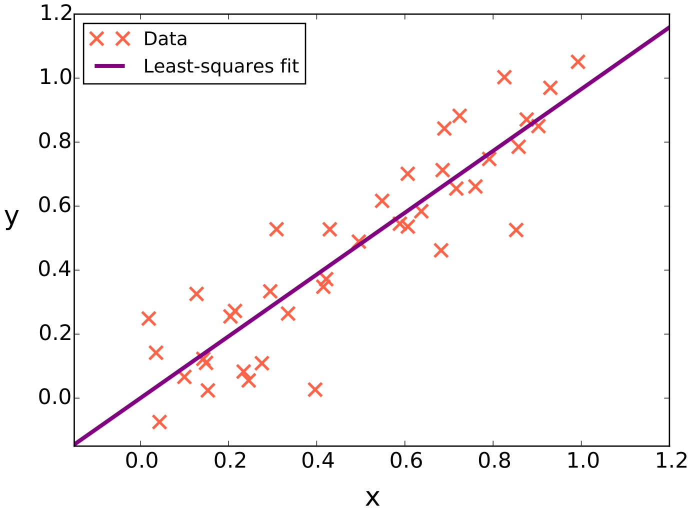

Figure 12.1: Linear model learnt via least-squares fitting for a simple example where there is just one feature ($p = 1$) and 40 examples ($n = 40$).

Example 12.2.1 (Linear model for GDP (continued))**.** The least-squares estimate for the regression coefficients in the linear GDP model is equal to

$$\vec{\beta}_{\text{LS}} = \begin{bmatrix} 1.019 \\ -0.111 \end{bmatrix}. \tag{12.13}$$

The GDP seems to be proportional to the population and inversely proportional to the unemployment rate. We now compare the fit provided by the linear model to the original data, as well as its prediction of the GDP of Tennessee:

$$\begin{array}{l c}
 & \begin{array}{cc} \text{GDP} & \text{Estimate} \end{array} \\
\begin{array}{l} \text{North Dakota} \\ \text{Alabama} \\ \text{Mississippi} \\ \text{Arkansas} \\ \text{Kansas} \\ \text{Georgia} \\ \text{Iowa} \\ \text{West Virginia} \\ \text{Kentucky} \\ \text{Tennessee} \end{array} &
\begin{pmatrix} 52\ 089 & 46\ 241 \\ 204\ 861 & 239\ 165 \\ 107\ 680 & 119\ 005 \\ 120\ 689 & 145\ 712 \\ 153\ 258 & 136\ 756 \\ 525\ 360 & 513\ 343 \\ 178\ 766 & 158\ 097 \\ 73\ 374 & 59\ 969 \\ 197\ 043 & 194\ 829 \\ 328\ 770 & 345\ 352 \end{pmatrix}
\end{array}$$

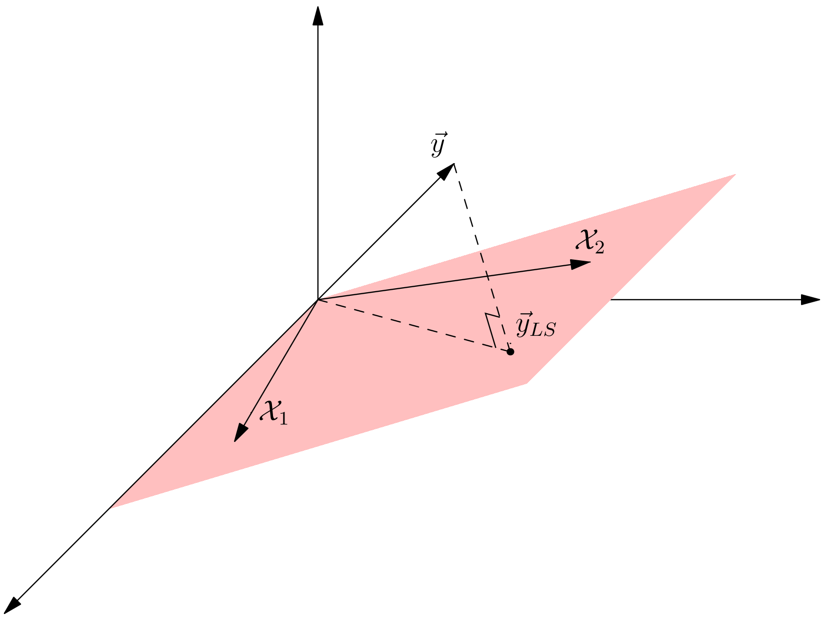

Figure 12.2: Illustration of Corollary 12.2.3. The least-squares solution is a projection of the data onto the subspace spanned by the columns of $\mathcal{X}$, denoted by $\mathcal{X}_1$ and $\mathcal{X}_2$.

△

12.2.1 Geometric interpretation

The following theorem, proved in Section 12.2.2, shows that the least-squares problem has a closed form solution.

Theorem 12.2.2 (Least-squares solution)**.** *For $p \geq n$, if $\mathcal{X}$ is full rank then the solution to the least-squares problem* (12.12) *is*

$$\vec{\beta}_{\text{LS}} := \left(\mathcal{X}^T \mathcal{X}\right)^{-1} \mathcal{X}^T \vec{y}. \tag{12.14}$$

A corollary to this result provides a geometric interpretation for the least-squares estimate of $\vec{y}$: it is obtained by projecting the response onto the column space of the matrix formed by the predictors.

Corollary 12.2.3. *For $p \geq n$, if $\mathcal{X}$ is full rank then $\mathcal{X}\vec{\beta}_{\text{LS}}$ is the projection of $\vec{y}$ onto the column space of $\mathcal{X}$.*

We provide a formal proof in Section 12.5.2 of the appendix, but the result is very intuitive. Any vector of the form $\mathcal{X}\vec{\beta}$ is in the span of the columns of $\mathcal{X}$. By definition, the least-squares estimate is the closest vector to $\vec{y}$ that can be represented in this way, so it is the projection of $\vec{y}$ onto the column space of $\mathcal{X}$. This is illustrated in Figure 12.2.

12.2.2 Probabilistic interpretation

If we model the noise in (12.5) as a realization from a random vector $\vec{Z}$ which has entries that are independent Gaussian random variables with mean zero and a certain variance σ^2, then we can interpret the least-squares estimate as a maximum-likelihood estimate. Under that assumption, the data are a realization of the random vector

$$\vec{Y} := \mathcal{X}\vec{\beta} + \vec{Z}, \tag{12.15}$$

which is an iid Gaussian random vector with mean $\mathcal{X}\vec{\beta}$ and covariance matrix $\sigma^2 I$. The joint pdf of $\vec{Y}$ is equal to

$$\begin{align} f_{\vec{Y}}\left(\vec{a}\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(\vec{a}_i - \left(\mathcal{X}\vec{\beta}\right)_i\right)^2\right) \tag{12.16}\\ &= \frac{1}{\sqrt{\left(2\pi\right)^n}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\left|\left|\vec{a} - \mathcal{X}\vec{\beta}\right|\right|_2^2\right). \tag{12.17}\end{align}$$

The likelihood is the probability density function of $\vec{Y}$ evaluated at the observed data $\vec{y}$ and interpreted as a function of the weight vector $\vec{\beta}$,

$$\mathcal{L}_{\vec{y}}\left(\vec{\beta}\right) = \frac{1}{\sqrt{\left(2\pi\right)^n}} \exp\left(-\frac{1}{2}\left|\left|\vec{y} - \mathcal{X}\vec{\beta}\right|\right|_2^2\right). \tag{12.18}$$

To find the ML estimate, we maximize the log likelihood. We conclude that it is given by the solution to the least-squares problem, since

$$\begin{align} \vec{\beta}_{\text{ML}} &= \arg\max_{\vec{\beta}} \mathcal{L}_{\vec{y}}\left(\vec{\beta}\right) \tag{12.19}\\ &= \arg\max_{\vec{\beta}} \log \mathcal{L}_{\vec{y}}\left(\vec{\beta}\right) \tag{12.20}\\ &= \arg\min_{\vec{\beta}} \left|\left|\vec{y} - \mathcal{X}\vec{\beta}\right|\right|_2^2 \tag{12.21}\\ &= \vec{\beta}_{\text{LS}}. \tag{12.22}\end{align}$$

12.3 Overfitting

Imagine that a friend tells you:

I found a cool way to predict the temperature in New York: It's just a linear combination of the temperature in every other state. I fit the model on data from the last month and a half and it's perfect!

Your friend is not lying, but the problem is that she is using a number of data points to fit the linear model that is roughly the same as the number of parameters. If $n \leq p$ we can find a $\vec{\beta}$ such that $\vec{y} = \mathcal{X}\vec{\beta}$ exactly, even if $\vec{y}$ and $\mathcal{X}$ have nothing to do with each other! This is called overfitting and is usually caused by using a model that is too flexible with respect to the number of data that are available.

To evaluate whether a model suffers from overfitting we separate the data into a training set and a test set. The training set is used to fit the model and the test set is used to evaluate the error. A model that overfits the training set will have very low error when evaluated on the training examples, but will not generalize well to the test examples.

Figure 12.3 shows the result of evaluating the training error and the test error of a linear model with $p = 50$ parameters fitted from n training examples. The training and test data are generated by fixing a vector of weights $\vec{\beta}^*$ and then computing

$$\vec{y}_{\text{train}} = \mathcal{X}_{\text{train}}\,\vec{\beta}^* + \vec{z}_{\text{train}}, \tag{12.23}$$

$$\vec{y}_{\text{test}} = \mathcal{X}_{\text{test}}\,\vec{\beta}^*, \tag{12.24}$$

where the entries of $\mathcal{X}_{\text{train}}$, $\mathcal{X}_{\text{test}}$, $\vec{z}_{\text{train}}$ and $\vec{\beta}^*$ are sampled independently at random from a Gaussian distribution with zero mean and unit variance. The training and test errors are defined as

$$\text{error}_{\text{train}} = \frac{\left|\left|\mathcal{X}_{\text{train}}\,\vec{\beta}_{\text{LS}} - \vec{y}_{\text{train}}\right|\right|_2}{\left|\left|\vec{y}_{\text{train}}\right|\right|_2}, \tag{12.25}$$

$$\text{error}_{\text{test}} = \frac{\left|\left|\mathcal{X}_{\text{test}}\,\vec{\beta}_{\text{LS}} - \vec{y}_{\text{test}}\right|\right|_2}{\left|\left|\vec{y}_{\text{test}}\right|\right|_2}. \tag{12.26}$$

Note that even the true $\vec{\beta}^*$ does not achieve zero training error because of the presence of the noise, but the test error is actually zero if we manage to estimate $\vec{\beta}^*$ exactly.

The training error of the linear model grows with n. This makes sense as the model has to fit more data using the same number of parameters. When n is close to $p := 50$, the fitted model is much better than the true model at replicating the training data (the error of the true model is shown in green). This is a sign of overfitting: the model is adapting to the noise and not learning the true linear structure. Indeed, in that regime the test error is extremely high. At larger n, the training error rises to the level achieved by the true linear model and the test error decreases, indicating that we are learning the underlying model.

12.4 Global warming

In this section we describe an application of linear regression to climate data. In particular, we analyze temperature data taken in a weather station in Oxford over 150 years.[1] Our objective is not to perform prediction, but rather to determine whether temperatures have risen or decreased during the last 150 years in Oxford.

In order to separate the temperature into different components that account for seasonal effects we use a simple linear model with three predictors and an intercept

$$\vec{y}_t \approx \vec{\beta}_0 + \vec{\beta}_1 \cos\left(\frac{2\pi t}{12}\right) + \vec{\beta}_2 \sin\left(\frac{2\pi t}{12}\right) + \vec{\beta}_3\, t \tag{12.27}$$

where $1 \leq t \leq n$ denotes the time in months (n equals 12 times 150). The corresponding matrix

[1] The data is available at http://www.metoffice.gov.uk/pub/data/weather/uk/climate/stationdata/oxforddata.txt.

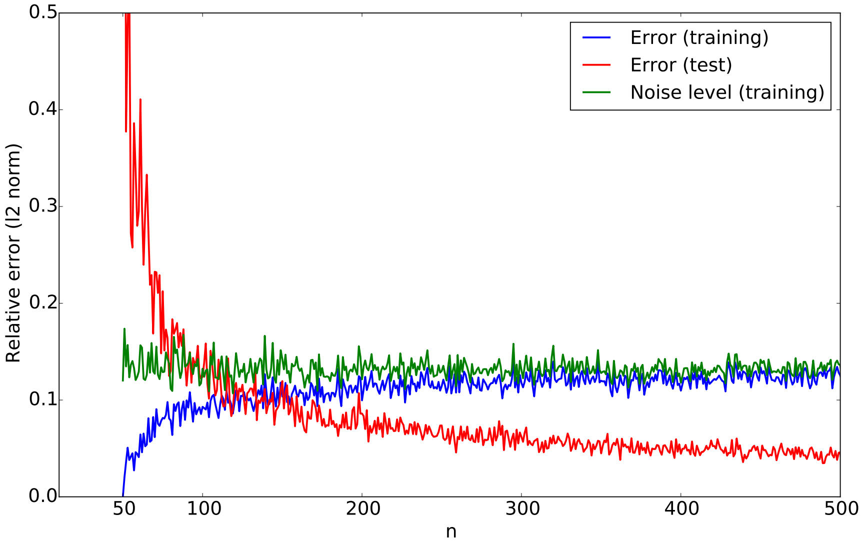

Figure 12.3: Relative ℓ_2-norm error in estimating the response achieved using least-squares regression for different values of n (the number of training data). The training error is plotted in blue, whereas the test error is plotted in red. The green line indicates the training error of the true model used to generate the data.

of predictors is

$$\mathcal{X} := \begin{bmatrix} 1 & \cos\left(\frac{2\pi t_1}{12}\right) & \sin\left(\frac{2\pi t_1}{12}\right) & t_1 \\ 1 & \cos\left(\frac{2\pi t_2}{12}\right) & \sin\left(\frac{2\pi t_2}{12}\right) & t_2 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & \cos\left(\frac{2\pi t_n}{12}\right) & \sin\left(\frac{2\pi t_n}{12}\right) & t_n \end{bmatrix}. \tag{12.28}$$

The intercept $\vec{\beta}_0$ represents the mean temperature, $\vec{\beta}_1$ and $\vec{\beta}_2$ account for periodic yearly fluctuations and $\vec{\beta}_3$ is the overall trend. If $\vec{\beta}_3$ is positive then the model indicates that temperatures are increasing, if it is negative then it indicates that temperatures are decreasing.

The results of fitting the linear model are shown in Figures 12.4 and 12.5. The fitted model indicates that both the maximum and minimum temperatures have an increasing trend of about 0.8 degrees Celsius (around 1.4 degrees Fahrenheit).

12.5 Proofs

12.5.1 Proof of Proposition 12.2.2

Let $\mathcal{X} = U\Sigma V_T$ be the singular-value decomposition (SVD) of $\mathcal{X}$. Under the conditions of the theorem, $\left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T y = V\Sigma U^T$. We begin by separating $\vec{y}$ into two components

$$y = UU^T y + \left(I - UU^T\right) y \tag{12.29}$$

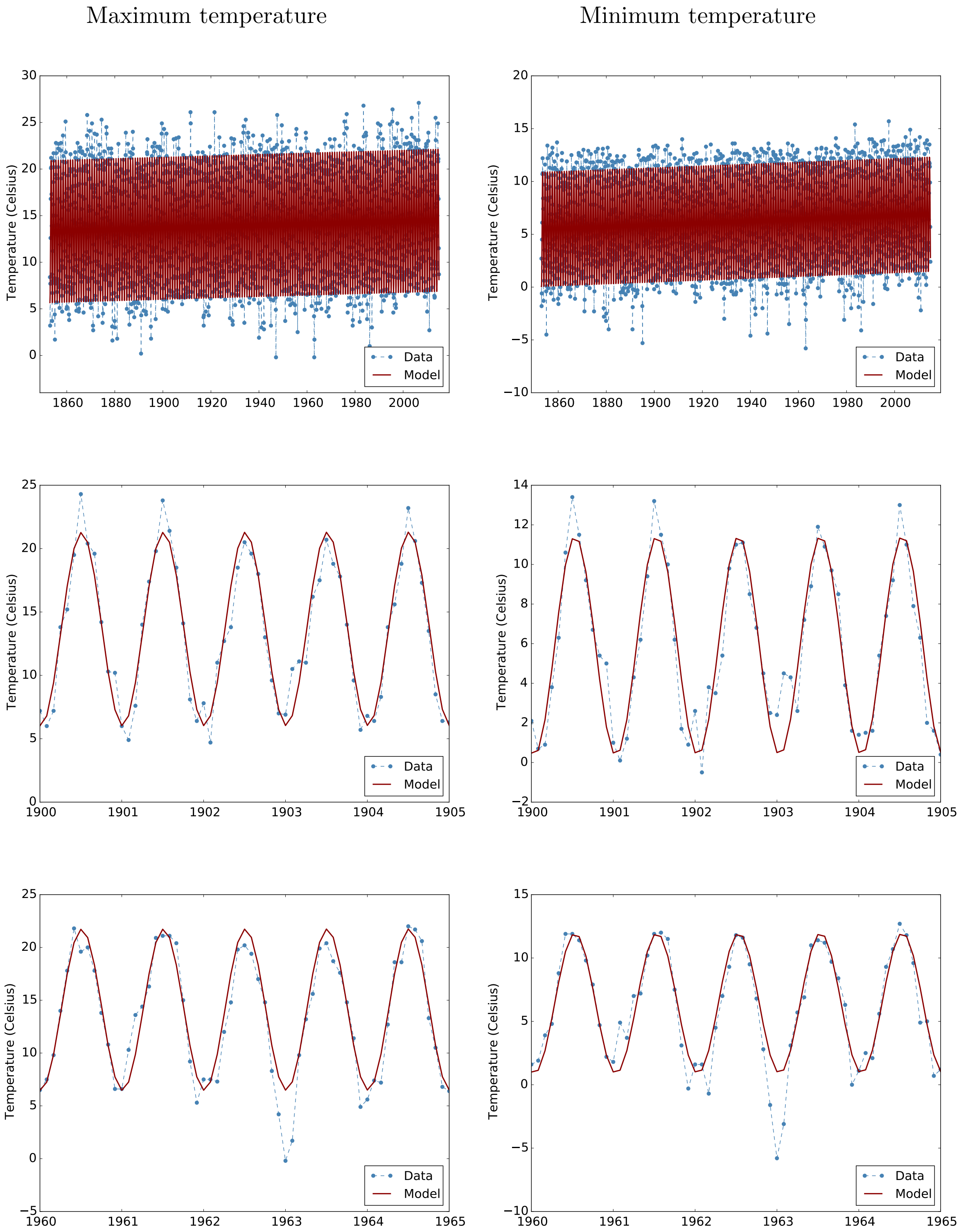

Figure 12.4: Temperature data together with the linear model described by (12.27) for both maximum and minimum temperatures.

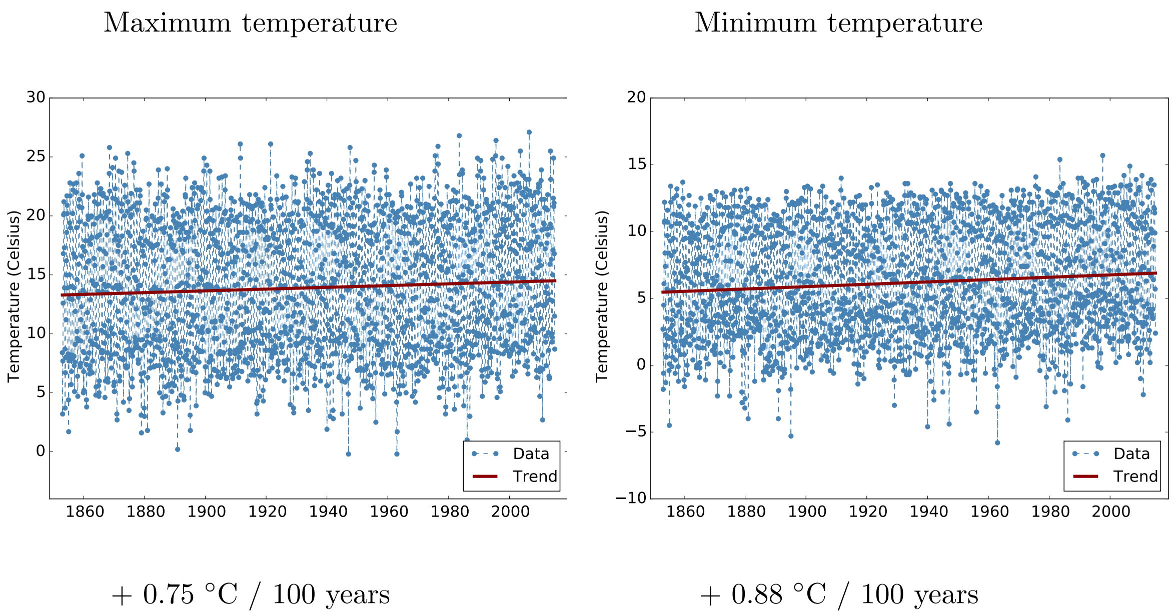

Figure 12.5: Temperature trend obtained by fitting the model described by (12.27) for both maximum and minimum temperatures.

where UU^Ty is the projection of $\vec{y}$ onto the column space of $\mathcal{X}$. Note that $\left(I-UU^T\right)y$ is orthogonal to the column space of $\mathcal{X}$ and consequently to both UU^Ty and $\mathcal{X}\vec{\beta}$ for any $\vec{\beta}$. By Pythagoras's Theorem

$$\left|\left|\vec{y}-\mathcal{X}\vec{\beta}\right|\right|_2^2 = \left|\left|\left(I-UU^T\right)y\right|\right|_2^2 + \left|\left|UU^Ty-\mathcal{X}\vec{\beta}\right|\right|_2^2. \tag{12.30}$$

The minimum value of this cost function that can be achieved by optimizing over $\tilde{beta}$ is $||\vec{y}_{\mathcal{X}^\perp}||_2^2$. This can be achieved by solving the system of equations

$$UU^Ty = \mathcal{X}\vec{\beta} = U\Sigma V_T\vec{\beta}. \tag{12.31}$$

Since $U^TU = I$ because $p \geq n$, multiplying both sides of the equality yields the equivalent system

$$U^Ty = \Sigma V_T\vec{\beta}. \tag{12.32}$$

Since $\mathcal{X}$ is full rank, Σ and V are square and invertible (and by definition of the SVD $V^{-1} = V^T$), so

$$\vec{\beta}_{\text{LS}} = V\Sigma U^Ty \tag{12.33}$$

is the unique solution to the system and consequently also of the least-squares problem.

12.5.2 Proof of Corollary 12.2.3

Let $\mathcal{X} = U\Sigma V^T$ be the singular-value decomposition of $\mathcal{X}$. Since $\mathcal{X}$ is full rank and $p \geq n$ we have $U^T U = I$, $V^T V = I$ and Σ is a square invertible matrix, which implies

$$\mathcal{X}\vec{\beta}_{\text{LS}} = \mathcal{X}\left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T y \tag{12.34}$$

$$= U\Sigma V^T \left(V\Sigma U^T U\Sigma V^T\right) V\Sigma U^T y \tag{12.35}$$

$$= UU^T y. \tag{12.36}$$