4 Linear models, least squares, and maximum likelihood

Here we shall learn the classical approach to fitting linear models via least squares and maximum likelihood, starting with the example of fitting a straight line in one dimension. We will see how this can easily be extended to higher dimensions and fitting nonlinear functions of the data. In the course of this we shall encounter the concepts of degrees of freedom, and the bias-variance decomposition. We will take the full Bayesian approach to this problem in chapter 9.

4.1 One-dimensional linear regression[1]

If we have two data points (x_1, y_1) and (x_2, y_2) then there is a unique straight line that fits these data. But what if we have three points or more, $\{x_i, y_i\}$?

There is no unique solution to this problem when stated in this way (unless the points are co-linear). We can imagine many different fits; but what is a good one?

Let the model for the straight line be $f(x;\theta)$, where θ denotes the parameters of the model, and $f_i = f(x_i;\theta)$ are the model predictions. A good fit in one sense will be a line for which the residuals of the fit, $\delta_i = y_i - f_i$, of the $i = 1 \ldots N$ points are small. We can combine these into a single metric by taking the sum of their squares (as squaring gets rid of the sign)

$$\mathrm{SS}_{\mathrm{res}} = \sum_i \delta_i^2 = \sum_i (y_i - f_i)^2. \tag{4.1}$$

This is the Euclidean distance between the data and the fit in an N-dimensional space, and is called the *sum of squared residuals* or the *residual sum of squares* (RSS). If we minimize this metric as a function of the parameters of the fit θ, then we can find the "best" solution. This is called *least squares minimization*. The idea of minimizing some error function (also called a cost, loss, or objective function) with respect to the model parameters is a key idea in fitting models. Mathematically we set

$$\frac{\partial \mathrm{SS}_{\mathrm{res}}}{\partial \theta_j} = 0 \tag{4.2}$$

for each of the model parameters θ_j, and then solve this set of equations for these parameters. We will have as many equations as parameters.

[1] "One-dimensional" here refers to the x being a one-dimensional quantity.

Suppose our model is $f = a + bx$, which is linear in the parameters (see section 4.2 for a definition). We want to find the best fitting coefficients, a and b.

$$\mathrm{SS}_{\mathrm{res}} = \sum_i (y_i - a - bx_i)^2. \tag{4.3}$$

Differentiating with respect to a and setting to zero gives

$$\sum_i -2(y_i - a - bx_i) = 0$$

$$\overline{y} - a - b\overline{x} = 0 \tag{4.4}$$

where $\overline{x} = \frac{1}{N}\sum_i x_i$ and $\overline{y} = \frac{1}{N}\sum_i y_i$. Differentiating with respect to b gives

$$\sum_i -2x_i(y_i - a - bx_i) = 0$$

$$\overline{xy} - a\overline{x} - b\overline{x^2} = 0. \tag{4.5}$$

Eliminating a from these two equations gives

$$b = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \tag{4.6}$$

$$= \frac{\widehat{\mathrm{Cov}}(x, y)}{\widehat{\mathrm{Var}}(x)} \tag{4.7}$$

where $\widehat{\mathrm{Cov}}(x, y)$ and $\widehat{\mathrm{Var}}(x)$ are the sample covariance and sample variance respectively (see sections 1.3 and 1.6.1).[2] We also get

$$a = \overline{y} - b\overline{x}. \tag{4.8}$$

The line goes through the mean of the data $(\overline{x}, \overline{y})$.

In the more general case of a linear model with J parameters, we will have a system of J linear equations that need to be solved (see sections 4.5 and 4.6).

4.1.1 Uncertainty estimates on the model parameters

Noting that

$$\overline{xy} - \overline{x}\,\overline{y} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \frac{\overline{x}}{N}\sum_{i=1}^{N} y_i$$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})y_i \tag{4.9}$$

we can write the estimate for the gradient (equation 4.6) as a weighted sum of the data points $\{y_i\}$

$$b = \frac{1}{NV_x}\sum_{i=1}^{N}(x_i - \overline{x})y_i \tag{4.10}$$

[2] We can write this in terms of sample estimates because a factor of $(N-1)/N$ appears in both the numerator and denominator, which then cancels.

where[3]

$$V_x = \overline{x^2} - \overline{x}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2. \tag{4.11}$$

Using the result in section 2.8.5, the variance of a linear combination of independently measured variables $\{y_i\}$ is

$$\mathrm{Var}(b) = \sum_i \left(\frac{\partial b}{\partial y_i}\right)^2 \sigma_i^2 \tag{4.12}$$

where σ_i is the standard deviation (uncertainty) in measurement y_i.[4] The x_i are assumed to be error free. From equation 4.10

$$\frac{\partial b}{\partial y_i} = \frac{x_i - \overline{x}}{NV_x} \tag{4.13}$$

so

$$\mathrm{Var}(b) = \sum_{i=1}^{N}\left(\frac{x_i - \overline{x}}{NV_x}\right)^2 \sigma_i^2. \tag{4.14}$$

If all data points have the same standard deviation, $\sigma_i^2 = \sigma^2$, we can write this (using also equation 4.11) as

$$\begin{aligned}\mathrm{Var}(b) &= \sigma^2 \sum_{i=1}^{N}\left(\frac{x_i - \overline{x}}{NV_x}\right)^2 \\ &= \sigma^2 \frac{NV_x}{(NV_x)^2} \\ &= \frac{\sigma^2}{NV_x}\,.\end{aligned} \tag{4.15}$$

You can likewise show that the variance in the intercept a is

$$\mathrm{Var}(a) = \frac{\sigma^2 \overline{x^2}}{NV_x} = \mathrm{Var}(b)\,\overline{x^2} \tag{4.16}$$

and the covariance between the gradient and intercept is

$$\mathrm{Cov}(a,b) = -\frac{\sigma^2 \overline{x}}{NV_x} = -\mathrm{Var}(b)\,\overline{x}. \tag{4.17}$$

Unless $\overline{x} = 0$, there is a non-zero covariance between the parameters. We interpret this as a parameter degeneracy: we can compensate for an increase in the gradient – to achieve almost as good a fit to the data – by decreasing the intercept. To remove this covariance we can translate our data to have $\overline{x} = 0$ prior to model fitting. The y-axis then goes through the mean of the data, so a small change in the gradient does not compensate for any change in the intercept.

[3] V_x is not quite the sample variance, because this would involve a factor of $N - 1$ rather than N.
[4] Equation 4.12 is exact because b is just a linear function of y_i, so the higher order terms in the Taylor expansion are zero.

The above analysis shows that we need to have uncertainty estimates in the y_i measurements to calculate the uncertainties in the parameters. What if we don't have any uncertainty estimates? In that case we must estimate them from the data. We could assume that the residuals $y_i - f_i$ have a common variance σ^2, which we estimate as $\sum_i (y_i - f_i)^2/(N-2)$. This estimate follows from the definition of variance in which the expected value of the residual is zero.[5] This is what the R function `lm` does (see section 4.1.5). Of course, if we *really* have zero uncertainty in the y_i measurements, then there can be no uncertainty in the parameters either, and the variance is zero (as follows from the above equations). If we have individual uncertainty estimates, then we might want to do a weighted fit (see section 4.1.4).

4.1.2 Uncertainties on predictions

Having found the function coefficients and their variances and covariance, we can predict the value of the function at any new point along with its uncertainty. If f_j is the prediction at x_j, then

$$f_j = a + bx_j \quad \text{and} \tag{4.18}$$

$$\mathrm{Var}(f_j) = \mathrm{Var}(a + bx_j). \tag{4.19}$$

It follows from the result in section 2.8.3 (equation 2.21), that

$$\mathrm{Var}(f_j) = \mathrm{Var}(a) + \mathrm{Var}(bx_j) + 2\,\mathrm{Cov}(a, bx_j). \tag{4.20}$$

Assuming that x_j has zero variance, then $\mathrm{Var}(bx_j) = x_j^2\,\mathrm{Var}(b)$. Using the results from the previous section, and writing $x_j = \overline{x} + \Delta x_j$, we see that

$$\begin{aligned}
\mathrm{Var}(f_j) &= \mathrm{Var}(a) + x_j^2\,\mathrm{Var}(b) + 2x_j\,\mathrm{Cov}(a,b) \\
&= \mathrm{Var}(b)\,(\overline{x^2} + x_j^2 - 2x_j\,\overline{x}) \\
&= \frac{\sigma^2}{NV_x}\left(\overline{x^2} - \overline{x}^2 + (\Delta x_j)^2\right) \\
&= \frac{\sigma^2}{N}\left(1 + \frac{(\Delta x_j)^2}{V_x}\right).
\end{aligned} \tag{4.21}$$

This is nonlinear in Δx_j. However, at points far from the original data $(\Delta x_j)^2/V_x \gg 1$, and so $\sqrt{\mathrm{Var}(f_j)} \sim |\Delta x_j|$, i.e. the standard deviation in the prediction grows linearly with the distance from the mean of the data.

4.1.3 Measures for goodness of fit

We have now fit the best straight line in a least squares sense, but is the fit any good? One measure of this is the quantity we minimized, the residual sum of squares of equation 4.1. Analogous to this we can define the *explained sum of squares* (also called the regression

[5] This is $N-2$ rather than N because we have estimated two model parameters from the data, and so have used up two degrees of freedom, a concept I will define later in section 4.3.

sum of squares), which measures how much the fitted values deviate from the mean of the data

$$\mathrm{SS}_{\mathrm{ess}} = \sum_i (f_i - \overline{y})^2. \tag{4.22}$$

The *total sum of squares* measures the total spread of the data about its mean,

$$\mathrm{SS}_{\mathrm{tot}} = \sum_i (y_i - \overline{y})^2. \tag{4.23}$$

It may come as no surprise that

$$\mathrm{SS}_{\mathrm{tot}} = \mathrm{SS}_{\mathrm{res}} + \mathrm{SS}_{\mathrm{ess}}. \tag{4.24}$$

The *residual standard error* reported by the `lm` function in R (we'll get to this in section 4.1.5) is

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{SS}_{\mathrm{res}}}{\nu}} \tag{4.25}$$

where ν is the degrees of freedom (section 4.3), which is $N - 2$ for this straight line fit.

Another useful metric is the *r squared* value – also known as the *coefficient of determination* – r^2. This is the fraction of the total sum of squares that is explained by the fit

$$r^2 = \frac{\mathrm{SS}_{\mathrm{ess}}}{\mathrm{SS}_{\mathrm{tot}}} = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}. \tag{4.26}$$

As the fit cannot explain more spread than is in the data, r^2 ranges from 0 to 1, with 0 being no fit and 1 a perfect fit (zero residuals). You can show that for a straight line fit this is equal to the square of the correlation coefficient.

Possibly more useful is the *adjusted r squared* value, defined as

$$r^2_{\mathrm{adj}} = 1 - \frac{\mathrm{RSE}^2}{\widehat{\mathrm{Var}}(y)} = 1 - \frac{\mathrm{SS}_{\mathrm{res}}/\nu}{\mathrm{SS}_{\mathrm{tot}}/(N-1)}. \tag{4.27}$$

This describes the degree of variance described by the fit compared to the sample variance in the data $\widehat{\mathrm{Var}}(y)$, which is the variance with no fit at all. Here ν is the number of degrees of freedom in $\mathrm{SS}_{\mathrm{res}}$ (i.e. after computing the fit) and $N - 1$ is the number of degrees of freedom in $\mathrm{SS}_{\mathrm{tot}}$ (i.e. after computing the average).

4.1.4 Linear model fitting with measurement errors

Sometimes we have data with individual error estimates σ_i on the measurements y_i. In that case we expect (and tolerate) deviations from the fit of this order for each point. The sensible thing to do is therefore to minimize

$$\chi^2 = \sum_i \frac{(y_i - f_i)^2}{\sigma_i^2} \tag{4.28}$$

instead of $\mathrm{SS}_{\mathrm{res}}$. (For now χ^2 is just a symbol. We'll learn about the χ^2 distribution in section 10.5.) Differentiating χ^2 with respect to b and setting to zero gives

$$\sum_i \frac{-2x_i(y_i - a - bx_i)}{\sigma_i^2} = 0. \tag{4.29}$$

Comparing this to equation 4.5, we see that the solution for the coefficients will have the same form as before, but with the averaged quantities replaced by their inverse variance weighted quantities. Working through the maths we get

$$b = \frac{(\sum \sigma_i^{-2})(\sum \sigma_i^{-2} x_i y_i) - (\sum \sigma_i^{-2} x_i)(\sum \sigma_i^{-2} y_i)}{(\sum \sigma_i^{-2})(\sum \sigma_i^{-2} x_i^2) - (\sum \sigma_i^{-2} x_i)^2} \tag{4.30}$$

and

$$a = \left(\frac{\sum \sigma_i^{-2} y_i}{\sum \sigma_i^{-2}}\right) - b\left(\frac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right) \tag{4.31}$$

where all the sums are of course over i. Comparing these to equations 4.6 and 4.8 we see that we have generalized the mean quantities involved to be the inverse variance weighted means. This follows because of the linearity of the model: we had essentially the same system of linear equations. The variances and covariance of the parameters are similarly modified by the inverse variance. We will see a generalized matrix formulation of this in section 4.5.1.

4.1.5 Using R: the `lm` function

R provides a flexible interface for doing linear modelling with `lm`. It uses a formula notation, so to fit a straight line model to $\{x_i, y_i\}$ we write `lm(y ~ x)`. The following shows what it provides.

```
x <- c(10.9, 12.4, 13.5, 14.6, 14.8, 15.6, 16.2, 17.5, 18.3, 18.6)
y <- c(24.8, 30.0, 31.0, 29.3, 35.9, 36.9, 42.5, 37.9, 38.9, 40.5)
plot(x, y)
model1 <- lm(y ~ x)
summary(model1)

  Call:
  lm(formula = y ~ x)

  Residuals:
      Min       1Q   Median       3Q      Max
  -4.2029 -1.3692 -0.6237   1.2761   5.8294

  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept)   4.5972     5.8853   0.781 0.457213
  x             1.9798     0.3815   5.190 0.000833 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 2.889 on 8 degrees of freedom
  Multiple R-squared:  0.771,Adjusted R-squared:  0.7424
```

```
  F-statistic: 26.93 on 1 and 8 DF,  p-value: 0.0008328

attributes(model1)
  $names
   [1] "coefficients"  "residuals"     "effects"       "rank"
   [5] "fitted.values" "assign"        "qr"            "df.residual"
   [9] "xlevels"       "call"          "terms"         "model"

  $class
  [1] "lm"
```

Under the heading `Residuals` we see the quartiles of the residuals about the fit. `Coefficients` shows the estimated values of the parameters and their standard deviations, computed as described above. The function also performs a t test to determine the significance of these parameters given their standard errors (the columns labelled `t value` and `Pr(>|t|)`). I will explain these in section 10.4 after we have learned about hypothesis testing. For now it is sufficient to consider smaller probabilities as suggesting that the parameter is more likely to be required in the fit. After this come several statistics, some of which I defined earlier in this chapter. The various quantities calculated by `lm` can be listed using `attributes` and can be accessed using the `$` symbol, e.g. with `model1$coefficients`. Regardless of any formal statistical tests, we should always inspect the quality of the fit by plotting it over the data, and by looking for any structure in the residuals. This is done by the following code, which produces figure 4.1.

R file: `linear_regression.R`

```
##### Demonstration of lm

x <- c(10.9, 12.4, 13.5, 14.6, 14.8, 15.6, 16.2, 17.5, 18.3, 18.6)
y <- c(24.8, 30.0, 31.0, 29.3, 35.9, 36.9, 42.5, 37.9, 38.9, 40.5)
model1 <- lm(y ~ x)
model1$fitted.values
predict(model1) # lm has a predict method
xv <- 10:20
#yv <- predict(model1, xv) # does not work: need to name the variable 'x'
yv <- predict(model1, data.frame(x=xv))
pdf("linear_regression.pdf", 8, 4)
par(mfrow=c(1,2), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
plot(x, y)
lines(xv, yv)
# lm has an abline method so you can instead do abline(model1)
plot(x, model1$residuals, ylab="residuals", pch=16)
abline(h=0, col="grey60")
dev.off()
```

In the above example we have no uncertainties on the y values. In this case, R just uses the residual standard error (equation 4.25) as the estimate of σ in equation 4.15, as explained earlier. In fact, this is the standard way to use `lm`. But what if we *do* have a value for σ? It seems that `lm` does not accommodate this, although we can easily write an R function to do so. If we have individual uncertainties on each data point, i.e. the σ_i are all different (heteroscedastic), then once again we can just use the equations above (such as equation 4.14 for the variance in b). However, we probably then want to give more weight

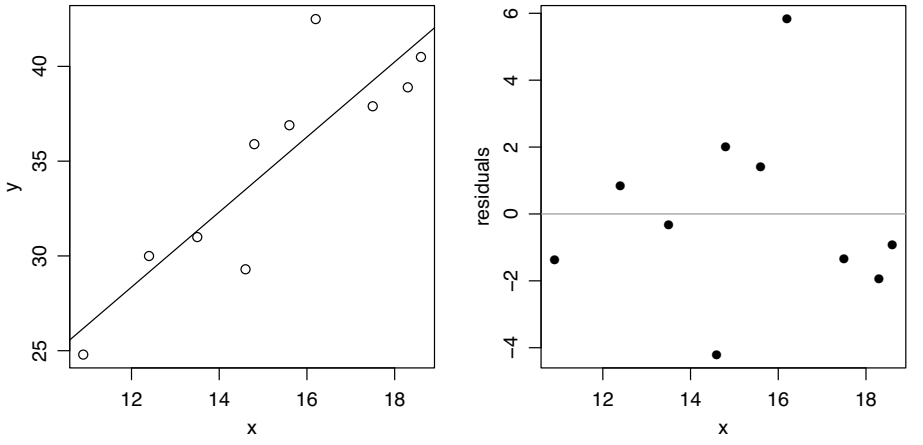

Fig. 4.1 Using the R function `lm` to do linear regression. Left shows the data and fit. Right shows the residuals of the fit.

in the fit to points with smaller uncertainties, and thus do a weighted least squares fit. We can do this with `lm` by giving it a vector of weights via the parameter `weights`. A common thing to do with data from a Gaussian measurement model is *inverse variance weighting*, where each weight is σ_i^{-2}, as described in section 4.1.4.

4.2 Linear models

The model $f = a + bx$ is called linear because it is linear in the parameters a and b. It does not need to be linear in the data x. Thus a model $f = \sum_j \beta_j x^j$ (for any range of j) is a linear model. If $h_j(x)$ is some arbitrary function of the data with *fixed* parameters, such as $\sin(\omega_j x)$, then $f = \sum_j \beta_j h_j(x)$ is also a linear model, because it is linear in the unknown (or "free") parameters $\{\beta_j\}$. A model is linear if the unknown parameters only appear to the first power, i.e. there are no terms like β_j^2 or $\beta_j \beta_{j'}$. A linear model $f(\theta)$, where θ are the unknown parameters and a is a constant, has the properties

$$f(\theta_1 + \theta_2) = f(\theta_1) + f(\theta_2) \quad \text{and} \tag{4.32}$$

$$f(a\theta) = af(\theta). \tag{4.33}$$

Linear regression means doing regression with a linear model. We shall see in section 4.6 how to do linear regression with nonlinear functions of the data. Nonlinear regression, in contrast, means the model is nonlinear in the parameters. We shall encounter this in chapter 9.

4.3 Degrees of freedom

The *degrees of freedom* of an estimate is the number of independent measures that are used to make that estimate. Imagine we have N independent measurements. This represents N pieces of information, so N degrees of freedom. Suppose we are now told the mean of the numbers. Given this, the numbers are no longer independent: once we know the first $N-1$ numbers, the final one is determined by the additional knowledge of the mean. By estimating one parameter from the data – here the mean – the remaining degrees of freedom goes down by one. If we now estimate the (sample) variance of the data, then the degrees of freedom that goes into this new calculation is $N-1$. This is because the mean is also used in the calculation, so the variance is essentially estimated from just $N-1$ independent numbers. The degrees of freedom in the estimates of the mean and variance (i.e. how many go into their calculation) are N and $N-1$ respectively. Each subsequent independent estimate from the data reduces the remaining degrees of freedom by one. Hence, if we fit a linear model with p parameters to N data points, the number of degrees of freedom in the data after having made the fit (or just "in the fit") is $N-p$. A quadratic ($p=3$) fit to seven data points, for example, would leave four degrees of freedom. We cannot fit a degree seven polynomial ($p=8$) to seven data points because the degrees of freedom would be negative; fit parameters would remain undetermined. Generally speaking, a more robust fit is one which has more degrees of freedom, e.g. a quadratic curve can normally be better fit to 20 than to seven data points.

This simple relation between the degrees of freedom and the number of parameters only applies when the model is linear in its parameters. The idea that the degrees of freedom is reduced by one per parameter does not apply to nonlinear models. This is because nonlinear models with very few free parameters can be made to fit almost any data arbitrarily well, depending on the model. Consider the function $y = a\sin(\omega x + \phi)$. With suitable choices of the three parameters a, ω, and ϕ, we can fit *any* data set with any number of data points *perfectly*. The function may look ridiculously implausible, but χ^2 (section 10.5) will be zero, and the idea that the degrees of freedom is $N-p$ is meaningless.

4.4 Maximum likelihood and maximum posterior probability

In the previous chapter we learned that the Bayesian approach to finding the parameters of a parametric model is to write down the likelihood and prior, the product of which is proportional to the posterior PDF over the parameters. The result is an entire probability distribution over the parameters given the data. (We will see in chapter 9 how to do this for the line fitting problem.) From this we could then define an estimator, such as the mean, median, or mode of the distribution. In this chapter, in contrast, we have solved for a single set of parameters by minimizing the sum of squared residuals. Although finding the full

posterior is normally preferable, there are situations were we are only interested in some "best" solution. We can define "best" as the maximum of the posterior and attempt to find this directly. If we want to use a uniform prior (and often we won't), then the maximum of the posterior coincides with the maximum of the likelihood. Finding this maximum with respect to the parameters is the procedure known as *maximum likelihood*.

We have a set of data $D = \{y_i\}$. The likelihood – probability density – of one data point for model M with parameters θ is $P(y_i|\theta, M)$. Assuming that the data have been measured independently, then from the rules of probability the likelihood of the whole data set is

$$P(\{y_i\}|\theta, M) = \prod_i P(y_i|\theta, M). \tag{4.34}$$

This is sometimes written as $L(\theta)$ (for brevity, or out of laziness), because we are interested in maximizing it with respect to the parameters for fixed data. But don't forget that it is still a PDF in the data! It is convenient to work with the log likelihood, as this converts the product into a sum

$$\ln L(\theta) = \ln \prod_i P(y_i|\theta, M) = \sum_i \ln P(y_i|\theta, M). \tag{4.35}$$

This is valid because the logarithm is a strictly monotonic function of its argument, so a maximum in $\ln L(\theta)$ is a maximum in $L(\theta)$. The maximum, which defines $\hat{\theta}$, is

$$\frac{d \ln L(\theta = \hat{\theta})}{d\theta} = 0. \tag{4.36}$$

If we're lucky this has a closed form and we can solve it analytically, as is the case with linear models. But often we have to solve numerically, something we will look at in section 12.6.

Having found the maximum likelihood solution, we still need to provide uncertainties; an estimate without an uncertainty is of limited value. One way to do this will be presented in section 7.1.

One of the positive features of maximum likelihood is that its solution is invariant under strictly monotonic transformations of the parameter, as we now see. Suppose the maximum of $L(\theta)$ is at $\hat{\theta}$. We transform the parameter via a strictly monotonic function $h(\theta)$. Such a transformation guarantees the function is invertible, i.e. $\theta(h)$ exists. The maximum likelihood solution is

$$0 = \frac{dL(\theta)}{d\theta} = \frac{L(\theta(h))}{dh} \frac{dh}{d\theta}. \tag{4.37}$$

As $dh/d\theta \neq 0$ (due to strict monotonicity), this equation is only satisfied when $dL(\theta(h))/dh = 0$, which gives a point estimate $\hat{h}$. But the above equation also tells us that this occurs at $\hat{\theta}$, so it follows that $\hat{h}(\theta) = h(\hat{\theta})$. Thus if we transform the parameter we can simply transform the maximum.

This invariance arises because the likelihood is not a *density* function in θ. Although we wrote the likelihood as $L(\theta(h))$ above, this is a shorthand for $P(D|\theta(h))$, a density function in D. Thus changing from θ to h does not change the density function. If we instead have some density function $P(\theta)$ in θ, then when we transform the parameter, the density function becomes $P(h) = P(\theta)d\theta/dh$. As $d\theta/dh$ is generally a function of θ,

a maximum in $P(\theta)$ does not correspond to a maximum in $P(h)$. Thus, in general, the maximum of a density function is not invariant under a transformation of its parameter (or parameters). We already saw an example of this on page 34, when we transformed the coordinates of a bivariate density function from Cartesian to radial coordinates.

Note that it is only the maximum of the likelihood which is invariant under strictly monotonic transformations of θ. The shape of the likelihood can change, so other statistics like the mean or the variance (used to characterize uncertainties) are generally not invariant.

4.4.1 Maximum likelihood estimate of the weighted mean of Gaussian variables

We have measured a set of values $\{y_i\}$ with corresponding uncertainties $\{\sigma_i\}$. Assume these uncertainties are Gaussian. What is the maximum likelihood estimate of the mean of the data?

The question tells us each point has been drawn from a Gaussian distribution with $\mathcal{N}(\mu, \sigma_i)$, where μ is the mean we wish to estimate. Thus the likelihood L is

$$L(\mu) = \prod_i \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{(y_i-\mu)^2}{2\sigma_i^2}\right] \tag{4.38}$$

$$\ln L(\mu) = -\sum_i \ln(\sigma_i\sqrt{2\pi}) - \sum_i \frac{(y_i-\mu)^2}{2\sigma_i^2}. \tag{4.39}$$

To get the maximum likelihood estimate we differentiate with respect to μ and set to zero

$$\begin{aligned} \sum_i \frac{y_i-\mu}{\sigma_i^2} &= 0 \\ \sum_i \frac{y_i}{\sigma_i^2} &= \mu \sum_i \frac{1}{\sigma_i^2} \\ \mu &= \frac{\sum_i y_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}. \end{aligned} \tag{4.40}$$

This justifies the use of an inverse variance weighted mean when we have errors on each data point (section 2.8.7): it is the maximum likelihood estimate. If the standard deviations are all the same, $\sigma_i = \sigma$, so we get $\mu = (1/N)\sum_i y_i$.

4.4.2 Least squares and χ^2 as a case of maximum likelihood

When doing linear regression (section 4.1) you may have been wondering why we do least *squares*. That is, why do we minimize $\sum_i (y_i - f_i)^2$? Why not minimize instead the sum of absolute deviations $\sum_i |y_i - f_i|$, for example? The answer is "maximum likelihood with Gaussian errors". The (log) likelihood (considered as a function of the model parameter μ) is given by equation 4.39, which we can write as

$$\ln L(\mu) = c - \frac{1}{2}\sum_i \frac{(y_i-\mu)^2}{\sigma_i^2} = c - \frac{1}{2}\chi^2 \tag{4.41}$$

for some constant c. Thus by minimizing the residual sum of squares (or χ^2 when we have errors on each point) we obtain the maximum likelihood solution. Gaussian error models are widespread, leading us to least squares.[6]

4.4.3 Minimizing L^1 and L^2 losses

The sum of squared residuals is also called the L^2 *loss function*, because it involves the residuals to the power of 2. (The L^2 *norm* or L^2 *distance* is the square root of this.) Suppose we have a normalized PDF $P(z)$. We would like to find the estimator $\hat{z}$ which minimizes the expected value of the squared residuals $E[(z-\hat{z})^2]$. From the definition of expectation (equation 1.23) this means that

$$\begin{aligned} 0 &= \frac{\partial}{\partial \hat{z}} \int (z-\hat{z})^2 \, P(z)\, dz \\ &= -2 \int (z-\hat{z})\, P(z)\, dz \\ \hat{z} &= \int z\, P(z)\, dz \;=\; E[z] \end{aligned} \tag{4.42}$$

as $\int P(z)dz = 1$. Thus provided $P(z)$ has finite zeroth, first, and second moments, the mean minimizes the L^2 loss function.

Which quantity minimizes the expected value of the absolute residuals $|z-\hat{z}|$ (the L^1 *loss function*)? We proceed in the same way. The function $|z-\hat{z}|$ looks like $\vee$, so its derivative with respect to $\hat{z}$ is -1 when $z-\hat{z}<0$ and $+1$ when $z-\hat{z}>0$. Hence we can write

$$\begin{aligned} 0 &= \frac{\partial}{\partial \hat{z}} \int |z-\hat{z}|\, P(z)\, dz \\ &= -\int_{-\infty}^{\hat{z}} P(z)\, dz + \int_{\hat{z}}^{+\infty} P(z)\, dz. \end{aligned} \tag{4.43}$$

The two integrals are equal when $\hat{z}$ is the median of $P(z)$ (by definition). Thus the median minimizes the L^1 loss function.

4.4.4 Maximum posterior

When we have a non-uniform prior $P(\theta)$, we cannot simply maximize the likelihood. We must instead maximize the posterior, the product of the likelihood with the prior. It's again easier to work with the logarithm. Bayes' theorem (equation 3.12; dependence on the model M is implicit) is

$$\ln P(\theta\,|\,D) \;=\; \ln L(\theta) + \ln P(\theta) - \ln Z(D) \tag{4.44}$$

[6] While it is sometimes fine to maximize the likelihood to find the best fitting parameters, we should not use this method to find the best model among a set of different models, e.g. to determine whether a polynomial or a sinusoidal model is the best curve. Why this is, and how we find the best model, will be explained in chapter 11.

where $Z(D)$ is the normalization constant (which is independent of θ, so will disappear when we differentiate). The maximum of this is often called the *maximum a posteriori* (MAP) estimate of the parameters. As the posterior is a density function in θ, the MAP is generally not invariant under a strictly monotonic transformation of θ (unlike the maximum likelihood, as we saw at the beginning of this section).

Of course, if the prior is uniform in θ, then the MAP will give the same solution as the maximum likelihood.[7] Yet once we introduce priors, the issue arises of how to represent the prior. If we say "uniform prior", do we mean uniform in θ or uniform in $\ln\theta$, for example? (Assigning priors is discussed in sections 3.5.5 and 5.3.) A prior uniform in θ is not uniform in $\ln\theta$. The shape of the prior is generally not invariant under transformations of the parameter ("reparametrizations"), and so the maximum of its product with the likelihood is also not invariant.

Likewise, other point estimates of the posterior as well as interval estimates – used to give uncertainties – are generally non-invariant under reparametrizations, so we usually will be concerned with the parameter representation. A general invariance is difficult to achieve (and may anyway not always be necessary). There are important types of invariance which can be attained, however, as we will see in section 5.3.

4.4.5 Maximum posterior estimate for Poisson data

We are given a set of N independent measurements $\{y_i\}$. We are told they have been drawn from a Poisson distribution with unknown parameter λ, but we do know a priori that $P(\lambda) \propto \exp(-\lambda/a)$ where a is known. What is the maximum posterior (MAP) estimate of λ?

It is instructive to first find the maximum likelihood solution. From the Poisson density function (equation 1.51) the likelihood is

$$L(\lambda) = \prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \tag{4.45}$$

the logarithm of which is

$$\ln L(\lambda) = \sum_{i=1}^{N} y_i \ln\lambda - \lambda - \ln(y_i!). \tag{4.46}$$

Differentiating with respect to λ and setting to zero we get the maximum likelihood solution,

$$\sum_{i=1}^{N}\left(\frac{y_i}{\lambda} - 1\right) = 0$$

$$\lambda = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{4.47}$$

[7] For this reason maximum likelihood is often thought of as a special case of MAP. This isn't quite correct though, because the concept of priors does not exist in the maximum likelihood approach; so the issue of how to represent the parameter doesn't arise.

Let's now find the MAP solution. The normalized prior is

$$P(\lambda) = \frac{1}{a} e^{-\lambda/a}, \tag{4.48}$$

from which we get

$$\frac{d \ln P(\lambda)}{d\lambda} = -\frac{1}{a}. \tag{4.49}$$

The log posterior (equation 4.44) is

$$\ln P(\lambda | \{y_i\}) = \ln L(\lambda) + \ln P(\lambda) + \text{constant} \tag{4.50}$$

where the constant is independent of λ. Differentiating with respect to λ and setting to zero gives

$$\sum_{i=1}^{N} \left(\frac{y_i}{\lambda} - 1 \right) - \frac{1}{a} = 0$$

$$\lambda = \frac{1}{N + 1/a} \sum_{i=1}^{N} y_i. \tag{4.51}$$

The effect of the prior is to reduce the estimate of λ relative to the maximum likelihood solution by an amount which depends on a. This is consistent with the prior, which favours smallers solutions. In the limit $a \to \infty$ the prior is flat (uniform improper prior), and the maximum posterior is just the maximum likelihood solution again.

4.5 Multi-dimensional linear regression

Let us now generalize linear regression to J dimensions. Borrowing language from machine learning, each "input" is a now a J-dimensional (column) vector $\mathbf{x}$, and the corresponding "output" is the scalar y. Given a set of $i = 1 \ldots N$ inputs and outputs, we want to fit a J-dimensional hyperplane[8] of y to $\mathbf{x}$. Our model for any one output y_i is

$$\begin{aligned} f_i &= \beta_0 + \sum_{j=1}^{J} x_{i,j} \beta_j \\ &= \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} \end{aligned} \tag{4.52}$$

where in the second line I have absorbed the constant number 1 into the vector to form the $J+1$ input vector $\mathbf{x}_i^{\mathsf{T}} = (1, x_{i,1}, \ldots, x_{i,j}, \ldots, x_{i,J})$. Here $\boldsymbol{\beta}$ is the $J+1$ dimensional parameter vector in which β_0 is the first element. We can write the set of N inputs as an

[8] The prefix "hyper" is used to indicate the higher dimensional generalization of a geometric concept, such as a plane, sphere, or volume.

$N \times (J+1)$ matrix X, in which each row is the vector of $J+1$ "features" for input i, i.e.

$$X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \vdots \\ \mathbf{x}_N^\mathsf{T} \end{pmatrix}. \tag{4.53}$$

This is called the *design matrix*. The model can now be written

$$\mathbf{f} = X\boldsymbol{\beta} \tag{4.54}$$

with the corresponding N outputs written as the vector $\mathbf{y}$.

Proceeding as with one-dimensional regression, we want to minimize

$$\begin{aligned} \mathrm{SS}_{\mathrm{res}} &= \sum_{i=1}^{N} (y_i - f_i)^2 \\ &= \sum_{i} (y_i - \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta})^2 \\ &= (\mathbf{y} - X\boldsymbol{\beta})^\mathsf{T}(\mathbf{y} - X\boldsymbol{\beta}) \end{aligned} \tag{4.55}$$

with respect to $\boldsymbol{\beta}$. This is quadratic in $\boldsymbol{\beta}$ so always has a turning point (and you can show that it is a minimum). Differentiating and setting to zero we get

$$\begin{aligned} 0 &= X^\mathsf{T}(\mathbf{y} - X\boldsymbol{\beta}) \\ X^\mathsf{T}X\boldsymbol{\beta} &= X^\mathsf{T}\mathbf{y} \\ \boldsymbol{\beta} &= (X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y} \end{aligned} \tag{4.56}$$

provided $X^\mathsf{T}X$ (the *information matrix*) is not singular. The model-predicted values of y are therefore

$$\begin{aligned} \mathbf{f} &= X\boldsymbol{\beta} \\ &= X(X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y} \end{aligned} \tag{4.57}$$

whereby the $N \times N$ matrix $X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$ is sometimes called the *hat matrix*, because it put a hat on – makes an estimator out of – $\mathbf{y}$.

Note that for a Gaussian likelihood, the log likelihood is $-\frac{1}{2}\mathrm{SS}_{\mathrm{res}}$, to within an additive constant. What is labelled above as the information matrix is a special case of a more general definition which will be encounter in section 5.3.2.

This method of getting a solution for the coefficients is known as *ordinary least squares*.

4.5.1 Generalized least squares

We can generalize line fitting further to the case of having different uncertainties on each of the N outputs $\{y_i\}$. We can even allow them to have correlations. Both of these are characterized by the $N \times N$ covariance matrix Σ, the elements of which are $c_{ij} = \mathrm{Cov}(y_i, y_j)$. The elements in the leading diagonal are the variances; the off-diagonal elements are the

covariances (see section 1.6.2). Assuming a Gaussian error model, the likelihood of these data is an N-dimensional Gaussian

$$\begin{aligned} L(\{y_i\}) &= \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\chi^2\right) \quad \text{where} \\ \chi^2 &= (\mathbf{y} - X\boldsymbol{\beta})^{\mathsf{T}}\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta}) \end{aligned} \tag{4.58}$$

(cf. equation 4.28). By minimizing χ^2 with respect to $\boldsymbol{\beta}$ we maximize the likelihood (assuming Σ is fixed). This is a generalization of ordinary least squares. The residual sum of squares metric in equation 4.55 is a special case in which Σ is the identity matrix. Differentiating χ^2 with respect to $\boldsymbol{\beta}$ and setting to zero gives the solution

$$\boldsymbol{\beta} = (X^{\mathsf{T}}\Sigma^{-1}X)^{-1}X^{\mathsf{T}}\Sigma^{-1}\mathbf{y} \,. \tag{4.59}$$

4.6 One-dimensional regression with nonlinear functions

As discussed in section 4.2, the "linear" in "linear regression" refers to linear in the parameters, not in the data. Provided the model is linear in its parameters, then the metric we are minimizing, $\mathrm{SS}_{\mathrm{res}}$, is a quadratic function of the parameters (so its derivative is linear in the parameters), yielding a unique minimum we can find analytically. We can use linear regression to fit models which are nonlinear in the data, by doing multi-dimensional regression in which the other dimensions are nonlinear expression of the data. One way to do this is with the polynomial expansion up to the Jth order

$$f_J(x) = \sum_{j=0}^{J} x^j \beta_j \,. \tag{4.60}$$

Expressed in this way, we can use all the machinery from the previous section.

The R code below uses `lm` to investigate fitting a nonlinear relationship. `model1` is the linear expression $y = \beta_0 + \beta_1 x$, `model2` is $y = \beta_0 + \beta_1 x + \beta_2 x^2$, and `model2b` drops the linear term, $y = \beta_0 + \beta_2 x^2$. `model3` is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$. The fits and residuals of the straight line and quadratic models are shown in figure 4.2. The code outputs are shown together with the code below. I suggest you work through it before reading on. The quadratic model gives a significantly better fit than the straight line one in terms of the residual standard error and the adjusted r-squared. The cubic model is only marginally better on this count. The t test (to be covered in chapter 10) suggests that neither the quadratic nor the cubic term is significant. This is not the case, however, because we see that the linear fit is much poorer than the quadratic fit. Both terms achieve low significances because this test only investigates what happens if we remove them one term at a time, in which case the other term fits much of the nonlinearity (this will be explained in section 10.4). The quadratic model without the linear term (`model2b`) fits the data much worse

than the full quadratic model, as suggested by the residual standard error (it's even worse than the straight line model). If you plot the fitted curve, it even seems to bend the wrong way. The reason is that a model $y = \beta_0 + \beta_2 x^2$ always has its turning point at $x = 0$.

R file: `nonlinear_functions.R`

```
##### Demonstration of linear regression with nonlinear functions of the
##### data using lm()

x <- c(0.9, 2.1, 2.9, 3.8, 5.3, 6.0, 7.0, 8.4, 9.1, 9.8, 11.1, 11.9, 13.2,
       13.8, 14.8)
y <- c(19.5, 13.8, 16.6, 11.6, 8.3, 9.9, 7.6, 6.6, 6.2, 5.4, 5.9, 5.5, 5.6,
       4.4, 4.7)
par(mfrow=c(2,2), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
plot(x,y) # looks nonlinear...

# Straight line model
model1 <- lm(y ~ x)
summary(model1) # r^2 is a bit low
abline(model1, col="blue", lw=2)
plot(x, model1$residuals, pch=20, col="blue", ylim=c(-5,5))
# we see structure in the residuals
abline(h=0)

# Quadratic model
model2 <- lm(y ~ x + I(x^2)) # note meaning of "+" and I()
summary(model2) # looks better than linear
xv  <- seq(from=0, to=30, by=0.1) # generate data for prediction,
                                  # as abline doesn't do curves
yv2 <- predict(model2, data.frame(x=xv))
plot(x,y)
lines(xv, yv2, col="red", lw=2)
plot(x, model2$residuals, pch=20, col="red", ylim=c(-5,5))
# we see less structure in residuals
abline(h=0)

# Cubic model
model3 <- lm(y ~ x + I(x^2) + I(x^3))
summary(model3)
# no evidence for cubic term; r^2 hardly drops compared to quadratic
yv3 <- predict(model3, data.frame(x=xv))
plot(x,y)
lines(xv, yv3, col="magenta", lw=2)
plot(x, model3$residuals, pch=20, col="magenta", ylim=c(-5,5))
# this looks no better than quadratic
abline(h=0)

# Quadratic model without the linear term
model2b <- lm(y ~ I(x^2)) # can we drop linear term in the quadratic model?
summary(model2b)
yv2b <- predict(model2b, data.frame(x=xv))
plot(x,y)
lines(xv, yv2b, col="brown", lw=2) # no we can't!
plot(x, model2b$residuals, pch=20, col="brown", ylim=c(-5,5))
abline(h=0)
```

```
# Plot linear and quadratic results together
pdf("nonlinear_functions.pdf", 8, 4)
par(mfrow=c(1,2), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
plot(x,y)
abline(model1, lw=2)
lines(xv, yv2, lw=2, lty=2)
plot(x,   model1$residuals, pch=20, ylim=c(-4.5,4.5), ylab="residuals")
points(x, model2$residuals, pch=4)
abline(h=0, col="grey60")
dev.off()
```

Here is an edited version of the summaries produced by R for each of the four models.

```
lm(formula = y ~ x)
Residuals:
    Min      1Q  Median      3Q     Max
-2.9938 -1.6293 -0.4736  1.3421  4.1089
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.2292     1.1652  13.929 3.43e-09 ***
x            -0.9312     0.1282  -7.264 6.33e-06 ***
Residual standard error: 2.136 on 13 degrees of freedom
Multiple R-squared:  0.8023,Adjusted R-squared:  0.7871

lm(formula = y ~ x + I(x^2))
Residuals:
     Min       1Q   Median       3Q      Max
-2.00968 -0.65973  0.04368  0.73096  2.41122
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.68163    1.08735  19.020 2.51e-10 ***
x           -2.53278    0.31632  -8.007 3.73e-06 ***
I(x^2)       0.10133    0.01946   5.207 0.000219 ***
Residual standard error: 1.231 on 12 degrees of freedom
Multiple R-squared:  0.9394,Adjusted R-squared:  0.9292

lm(formula = y ~ x + I(x^2) + I(x^3))
Residuals:
     Min       1Q   Median       3Q      Max
-2.02586 -0.50426 -0.05821  0.32114  2.68788
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.230245   1.602753  13.870 2.59e-08 ***
x           -3.566737   0.860392  -4.145  0.00163 **
I(x^2)       0.260299   0.124959   2.083  0.06137 .
I(x^3)      -0.006712   0.005215  -1.287  0.22451

Residual standard error: 1.199 on 11 degrees of freedom
Multiple R-squared:  0.9473,Adjusted R-squared:  0.9329

lm(formula = y ~ I(x^2))
Residuals:
    Min      1Q  Median      3Q     Max
-3.2085 -2.6313 -0.5933  1.2637  6.6229
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.91773    1.19060   10.85 6.93e-08 ***
```

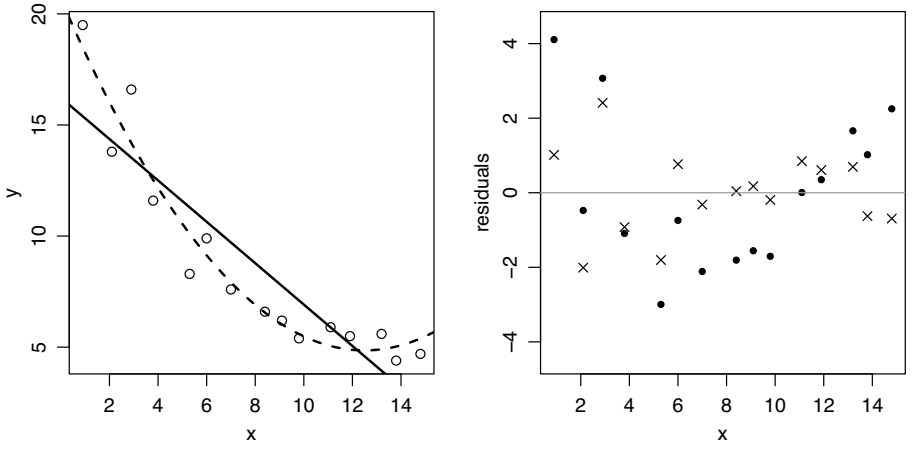

Fig. 4.2 Demonstration of linear regression with nonlinear functions of the data. Left shows the data (open circles) and fit for the straight line model (solid line) and quadratic model (dashed line). Right shows the residuals for the straight line model (filled circles) and quadratic model (crosses).

```
I(x^2)       -0.05017    0.01100   -4.56 0.000535 ***
Residual standard error: 2.979 on 13 degrees of freedom
Multiple R-squared:  0.6153,Adjusted R-squared:  0.5858
```

Using linear least squares on expansions of the data – so-called *basis functions* or *basis expansions* – is a powerful approach to defining and fitting nonlinear functions of data. We will look at these further in section 12.3.

4.7 Least squares with errors on both axes

So far we have considered the x variable (or $\mathbf{x}$ variables) as the "independent" variable and y as the "dependent" variable, because we have been fitting a model $f(x)$ or $f(\mathbf{x})$. This is because we think of x as being fixed, i.e. noise free. If y were instead noise free, then there is nothing to stop us from fitting a model $x = f^{-1}(y)$ instead. Least squares does not generally give the same solution – line in (x, y) space – because we are minimizing a different sum of squares.

But what do we do if we have errors in x as well as y? Errors in x must affect the solution. This is a so-called *errors in variables* problem. Let x' and y' represent the model-predicted (i.e. noise-free) values of x and y respectively. Our generative model is thus

$$y' = a + bx'. \tag{4.61}$$

Suppose we have the measurements $\{x_i\}$ and $\{y_i\}$ with corresponding uncertainties (Gaussian standard deviations) $\{\varsigma_i\}$ and $\{\sigma_i\}$ respectively. We now consider the residuals in both

variables, $x - x'$ and $y - y'$. A plausible generalization of the χ^2 metric (equation 4.28) is the sum of squares of error-weighted residuals (Deming, 1943),

$$\mathrm{SS}_{\mathrm{res}} = \sum_i \left(\frac{y_i - y'_i}{\sigma_i}\right)^2 + \left(\frac{x_i - x'_i}{\varsigma_i}\right)^2, \tag{4.62}$$

as this now includes the residuals in x. If $\varsigma_i = \varsigma$ and $\sigma_i = \sigma$ for all i, then this can be written as

$$\mathrm{SS}_{\mathrm{res}} = \frac{1}{\sigma^2}\sum_i (y_i - a - bx'_i)^2 + \lambda^2(x_i - x'_i)^2 \tag{4.63}$$

where $\lambda = \sigma/\varsigma$ is the ratio of the standard deviations. This has a (slightly complicated) analytic solution – for a, b, and the $\{x'_i\}$ – which depends not on the individual standard deviations, but only on λ. When $\lambda = 1$ the method is sometimes called *orthogonal least squares*, because it corresponds to measuring the residuals of the fit not parallel to the y-axis, but orthogonal to the line being fit.

This approach can be generalized for the case of non-zero covariance between the x and y variables. Let $\mathbf{z}_i = (x_i, y_i)^\mathsf{T}$ and $\mathbf{z}'_i = (x'_i, y'_i)^\mathsf{T}$ be the measured and true (unknown) positions of point i, respectively, and let their covariance be

$$\Sigma_i = \begin{bmatrix} \sigma^2_{x,i} & \rho\sigma_{x,i}\sigma_{y,i} \\ \rho\sigma_{x,i}\sigma_{y,i} & \sigma^2_{y,i} \end{bmatrix}. \tag{4.64}$$

Adopting a Gaussian measurement model, the likelihood for this single data point is a bivariate Gaussian (see equation 1.69)

$$L(a, b) = \frac{1}{2\pi|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z}_i - \mathbf{z}'_i)^\mathsf{T}\Sigma_i^{-1}(\mathbf{z}_i - \mathbf{z}'_i)\right). \tag{4.65}$$

Solving this for the two model parameters is known as *total least squares*, and in general requires a numerical solution. We will see in section 9.4 how this can be generalized further for the case of arbitrary error models.

4.8 Bias-variance decomposition

I finish up this chapter with a useful way of looking at the sum of squared residuals.

Suppose we have a set of data $\{x\}$ (they could be scalar or vector) and corresponding values $\{y\}$, for which we assume a relationship

$$y = f(x) + \epsilon \tag{4.66}$$

exists, where $f(x)$ is an unknown deterministic function, and ϵ is a noise source with mean zero and variance σ^2. Fitting a model to these data means finding a good estimator for $f(x)$, which I label $\hat{f}(x)$. The best estimator in the least squares sense is the one which minimizes the sum of squared residuals $\sum_i (y_i - \hat{f}(x_i))^2$. I will now show that this metric can be written in terms of the bias and the variance of the estimator. First, because f is a deterministic function and the noise is additive, $E[f] = f$. Second, since $E[\epsilon] = 0$, it

follows from the previous point and equation 4.66 that $E[y] = f$. Third, we can use the definition of variance (equation 1.30) to write

$$E[\hat{f}^2] = \mathrm{Var}(\hat{f}) + E[\hat{f}]^2 \quad \text{and} \tag{4.67}$$
$$E[y^2] = \mathrm{Var}(y) + E[y]^2 = \mathrm{Var}(y) + f^2. \tag{4.68}$$

Using these we see that the expected squared residual is

$$\begin{aligned} E[(y-\hat{f})^2] &= E[y^2] + E[\hat{f}^2] - 2E[y\hat{f}] \\ &= \mathrm{Var}(y) + f^2 + \mathrm{Var}(\hat{f}) + E[\hat{f}]^2 - 2fE[\hat{f}] \\ &= \mathrm{Var}(y) + \mathrm{Var}(\hat{f}) + (E[\hat{f}] - f)^2 \\ &= \mathrm{Var}(y) + \mathrm{Var}(\hat{f}) + E[\hat{f} - f]^2 \\ &= \sigma^2 + \mathrm{Var}(\hat{f}) + \mathrm{bias}^2. \end{aligned} \tag{4.69}$$

The bias is defined as $E[\hat{f} - f]$, the expected difference between the estimator and the true function. Hence the expected squared residual is the sum of the estimator's variance and the square of the estimator's bias, plus an irreducible variance σ^2 due to the noise in the true relationship. A perfect model would have zero bias and zero variance. This is usually unachievable. Good models can be obtained either by reducing their bias, or their variance, or both. When we fit a model in practice, we usually have to trade off the bias against the variance: decreasing one often increases the other. We shall see this in chapter 12.

As we do not know the true model, $f(x)$, we cannot calculate this bias and this variance. The above is a way of thinking about the properties of the estimator (and perhaps evaluating them with simulations). But we can compute similar things from real data. Let $\delta_i = y_i - \hat{f}_i$ for $i = 1 \ldots N$ be the residuals of a model fit. The root mean square (RMS) of these is

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}\,. \tag{4.70}$$

Contrast this with their standard deviation (SD) (neglecting the $N/(N-1)$ bias correction factor)

$$\mathrm{SD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\delta_i - \overline{\delta})^2} \tag{4.71}$$

in which

$$\overline{\delta} = \frac{1}{N}\sum_{i=1}^{N}\delta_i \tag{4.72}$$

is a measure of the bias. Whereas RMS measures the total spread of the residuals including

the bias, SD measures their spread about the bias. We see that

$$\begin{aligned} \mathrm{SD}^2 &= \frac{1}{N}\sum_{i=1}^{N}(\delta_i^2 + \bar{\delta}^2 - 2\bar{\delta}\delta_i) \\ &= \mathrm{RMS}^2 + \bar{\delta}^2 - 2\bar{\delta}^2 \\ \mathrm{RMS}^2 &= \mathrm{SD}^2 + \bar{\delta}^2 \end{aligned} \tag{4.73}$$

which is analogous to equation 4.69, except that now the irreducible variance is part of the bias and standard deviation. It is possible to have estimators which have a small scatter in their residuals, i.e. a small SD, but still a large RMS due to a large bias.