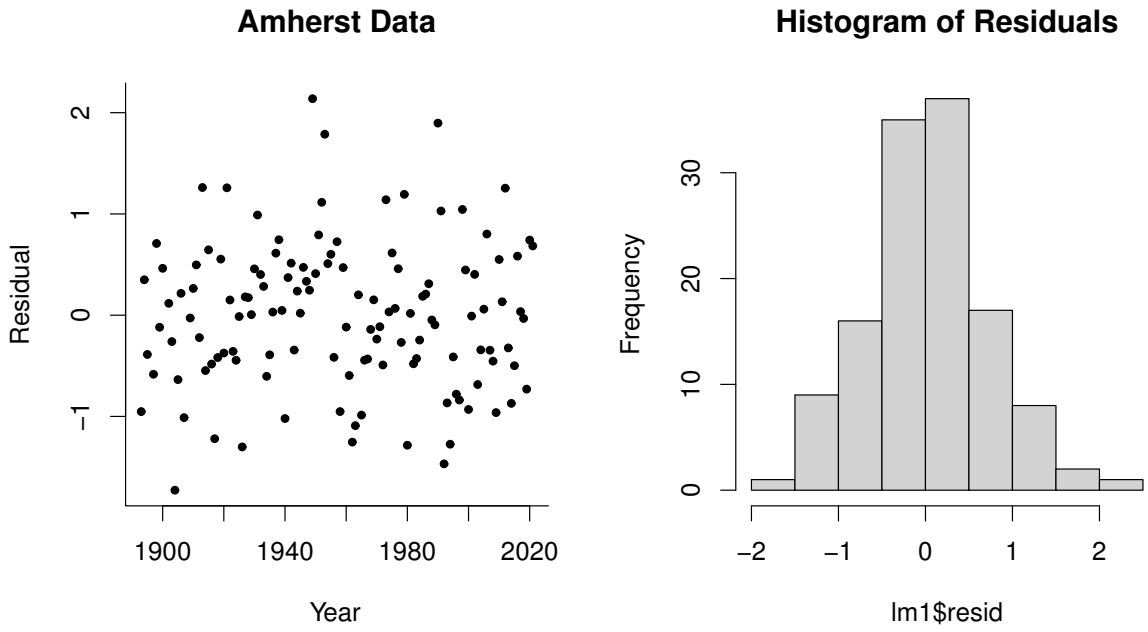

Figure 2.7 *Plots of residuals from a linear model fitted to Amherst data. (a) Residual versus x value. (b) Histogram of residuals.*

question the normality of the distribution. In fact, however, we shall show that formal tests of fit do not contradict normality in this case.

2.9.2 Tests of normality: Shapiro-Wilk and its relatives

The graphical techniques just described allow one to make a visual assessment of whether or not the data are normally distributed, and in many cases, such an assessment is all that is needed — for example, if the plots allow us to identify specific observations which are inconsistent with the rest of the data, then there is no need to make a formal test out of it. However it may also be the case, especially when inspection of the probability plot reveals nothing obviously wrong with the model, that one wants to follow this up with a test, which leads to a formal decision whether to accept or reject the normal distribution. A number of different methods of doing this have been proposed. In this section, we study methods based on probability plots; alternative methods based on the *empirical distribution function* are discussed in the next subsection.

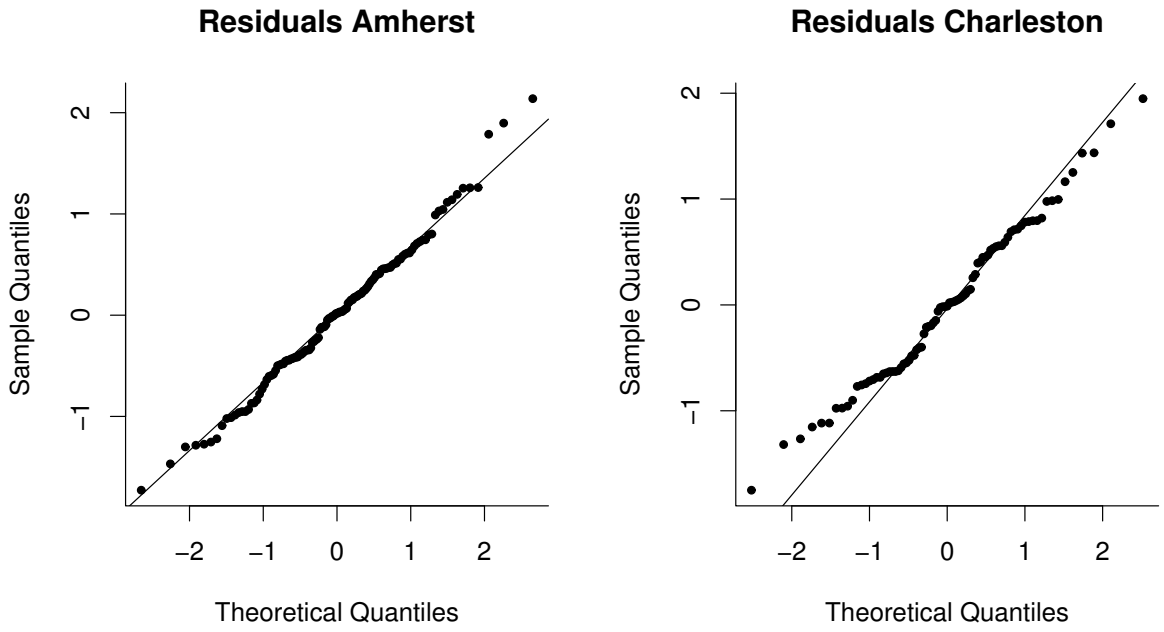

Figure 2.8 *Normal probability plot of residuals from (a) Amherst data, (b) Mount Airy—Charleston data.*

A probability plot, such as one of those in Figure 2.8, can itself be thought of as a regression experiment in which we try to measure how close the relationship between observed and expected values is to a straight line. For at least the first part of this discussion, we ignore the covariates in the linear model and assume $e_i = y_i - \bar{y}$, where $y_1, \ldots, y_n$ are independent normal random variables with common mean μ and variance σ^2, and $\bar{y}$ is the sample mean.

One of the most famous and best tests for normality is the Shapiro-Wilk test, introduced by Shapiro and Wilk [61]. This test relies on trying to construct a linear regression of e_i^* on the ith expected standard normal order statistic m_i for $1 \leq i \leq n$. This is complicated, because not only are the exact values of m_i rather hard to compute, but in the original version of the test, it was also necessary to know the variances and covariances of the e_i^* under the null hypothesis of a normal distribution. This is because the test used *generalized least squares* (GLS) to compute the best straight line for e_i^* against m_i. (See Chapter X.X for a description of GLS for a general

regression problem.) The result is a test statistic of the form

$$W = \frac{(\sum a_i e_i^*)^2}{\sum e_i^{*2}}.$$

Tables for the coefficients a_i and the percentage points of W were given by Shapiro and Wilk for samples up to size 50.

Subsequent authors have found ways of simplifying the method without sacrificing much power. Shapiro and Francia [60] simplified the computation of the test statistic by simply using the squared correlation coefficient between e_i^* and m_i,

$$W' = \frac{(\sum m_i e_i^*)^2}{\sum m_i^2 \sum e_i^{*2}},$$

and they gave tables of the distribution of W' for $n = 35$ and for values of n between 50 and 100. However, even this method assumes that one knows the exact value of m_i.

Looney and Gulledge [37] simplified the test further by considering various approximations to m_i. Two approximations in common use are $z_i^* = z_{(i-0.5)/n}$ or $z_i^\dagger = z_{(i-0.375)/(n+0.25)}$, which was originally proposed by Blom [8] and is generally considered superior to z_i^*. Rather than W_i', they gave the result directly in terms of the correlation coefficient

$$r^* = \frac{\sum z_i^\dagger e_i^*}{\sqrt{\sum z_i^{\dagger 2} \sum e_i^{*2}}},$$

giving tables for the significant values of r^* (small values of r^* are significant, because in a perfect probability plot, r^* would be very close to 1, the maximum value possible).

A final comment is that *none* of these methods really works correctly in a regression problem, because in regression the residuals do not all have the same variance and are correlated, in contrast with the simple case of independent observations from a common normal distribution. We shall go into these issues in detail in Chapter X.X, but for now, it is sufficient to note that they can cause problems.

In view of these last comments, the best recommendation is to use the correlation statistic r^*, but if time and computing resources permit, to compute the percentage points and p-values by simulation. The simulation steps proceed as follows:

1. Select the number of simulations to be used, M say.
2. For the I'th simulation ($1 \le I \le M$), generate n independent $N(0,1)$ observations using a standard random number generator. These will be the y values in the simulated experiment. The test statistic we are computing is invariant to changes in the true values of the regression coefficients, and to the variance of the y values, so for the purpose of the simulation, it is sufficient to assume that all the y values have mean 0 and variance 1.
3. Carry out a linear regression of $y_1, ..., y_n$ on $x_{ij},\ 1 \le i \le n,\ 1 \le j \le p$, the same x values as in the original experiment, and compute the residuals e_i.

4. Rearrange the residuals e_i in increasing order, and compute the r^* statistic — we write this r_I^* to indicate that it is the I'th simulated value of r^*.
5. When steps 2–4 have been completed M times, rearrange all the r_I^* values in increasing order and use these to form empirical percentage points for the distribution. Alternatively, compute

$$p = \frac{\text{Number of simulations for which } r_I^* < r^*}{M},$$

where r^* is the test statistic from the original experiment. This gives an approximate p-value for the test. Values of $p < .05$ create doubts about the suitability of the normal distribution; $p < .01$ creates more conclusive evidence that the normal distribution is not suitable for this data set.

One further comment here is that there may be some advantage in using either the standardized or the studentized (Chapter 3) residuals in place of the simple residuals.

2.9.3 Tests of normality based on the EDF

A second very widely-used class of tests of a normal distribution (or of any other specified distributional family) are those based on the empirical distribution function (EDF). Particular tests in this category include the Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling tests.

Suppose we have n observations which are independent and identically distributed (i.i.d.) with some distribution function $F(y)$ on $-\infty < y < \infty$. Note that we are not necessarily assuming that the range of the distribution covers the whole of the interval $(-\infty, \infty)$, but it is convenient to assume that F is defined everywhere, setting $F(y) = 0$ if y is below the left-hand endpoint and $F(y) = 1$ if y is above the right-hand endpoint of the distribution. In that case, the regions on which $F(y)$ is either 0 or 1 will play no role in the test statistic computations to follow. The *empirical distribution function* is the function

$$F_n(y) = \frac{1}{n}\{\text{ Number of observations } \leq y\},$$

defined for each $y \in (-\infty, \infty)$. The idea of EDF tests is that if we want to test a null hypothesis $F = F_0$, where F_0 is some given distribution function, we construct a test based on some measure of "closeness" between F_n and F_0.

Three commonly used such measures are the *Kolmogorov-Smirnov* statistic,

$$D = \max_{-\infty < y < \infty} |F_n(y) - F_0(y)|,$$

the *Cramér-von Mises* statistic,

$$W^2 = n\int_{-\infty}^{\infty} \{F_n(y) - F_0(y)\}^2 dF_0(y),$$

and the *Anderson-Darling* statistic,

$$A^2 = n\int_{-\infty}^{\infty} \frac{\{F_n(y) - F_0(y)\}^2}{F_0(y)\{1 - F_0(y)\}} dF_0(y).$$

The Kolmogorov-Smirnov is probably the most widely-known and widely-used of the three test statistics, though it is not necessarily the best in terms of power. Comparing the other two, the Anderson-Darling test puts more weight on the tails of the distribution, and therefore is often recommended for use when one is particularly concerned about departures from F_0 in the tails.

In practice, there are computing formulas which simplify the computation of these test statistics. If the original data are ordered and have the values $y_1^* \leq ... \leq y_n^*$, then we define u_i^* for $1 \leq i \leq n$ by $u_i^* = F_0(y_i^*)$, and let

$$D^+ = \max_i \left(\frac{i}{n} - u_i^* \right), \quad D^- = \max_i \left(u_i^* - \frac{i-1}{n} \right).$$

We then have

$$\begin{aligned} D &= \max(D^+, D^-), \\ W^2 &= \sum_i \left(u_i^* - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}, \\ A^2 &= -n - \frac{1}{n} \sum_i \{ (2i-1) \log u_i^* + (2n+1-2i) \log(1-u_i^*) \}. \end{aligned}$$

To define percentage points for these tests, the first thing to note is that they are *distribution-free* in the following sense: if F_0 is a continuous c.d.f. and is completely specified (no unknown parameters), then the distribution of D, W^2 and A^2 does not depend on F_0. The reason is that when F_0 is the true c.d.f, the transformation from y_i^* to u_i^* is a probability integral transformation to the uniform distribution, so that we are in effect testing that $u_1^*, ..., u_n^*$ are the order statistics from a uniform distribution on $[0,1]$. This is convenient, because it means that percentage points and p-values may be calculated independently of the true F_0.

In practice, however, things are not as simple as that. Usually F_0 is not completely specified, but only specified up to some unknown parameters. An example is when we are testing that F_0 is the normal distribution with mean μ and variance σ^2. In that case, μ and σ^2 are usually not known in advance, but are estimated using the sample mean and the sample standard variance. In that case, however, the calculations of percentage points and p-values must take the estimation of μ and σ^2 into account. Some tables and computer packages do this — in other words, quoted percentage points and p-values allow for the fact that μ and σ^2 are estimated.

In principle, when these tests are applied to residuals from a regression equation, the computation of percentage points should take into account that they are residuals from a regression, and should not merely treat them as independent observations from a normal distribution. In practice, this point is usually ignored. Tables and computer packages allow for the estimation of the mean and variance but not for the full joint distribution of the residuals. It is possible to allow for this, however, with a simulation procedure similar to that outlined at the end of Section 2.9.2. We expand on this further in the next section.

EDF tests are described in considerable detail in the book by D'Agostino and Stephens [21], which we would recommend for further reading.

2.9.4 Implementation in R

Two goodness of fit tests are built into base R: `shapiro.test` implements the Shapiro-Wilk test, and `ks.test` applies the Kolmogorov-Smirnov test.

As an example, let's first fit the linear regression to the Amherst dataset, using

```
lm1=lm(Temp~Year,data='Amherst')
shapiro.test(lm1$resid)
ks.test(rstandard(lm1),y='pnorm')
```

produces the outputs

```
Shapiro-Wilk normality test

data:  lm1$resid
W = 0.99136, p-value = 0.6259

One-sample Kolmogorov-Smirnov test

data:  rstandard(lm1)
D = 0.042636, p-value = 0.976
alternative hypothesis: two-sided
```

The Shapiro-Wilk statistic is applied directly to the residuals from the linear model, and indicates a p-value of 0.6259. Since this is well above 0.05, we conclude that the assumption of normally distributed errors is not rejected by this test.

Two points should be made about the Kolmogorov-Smirnov test in this context. The first is the condition `y='pnorm'` — we have to specify the distribution being tested, and in this case it is `pnorm`, that is, the standard normal distribution. (We could test other distributions with different specifications, for example, `y='punif'` would test the null hypothesis that the data are uniformly distributed on $(0, 1)$.) The second point follows from the first however — `pnorm` refers specifically to the normal distribution with mean 0 and variance 1, so `ks.test` would not give the correct answer if the distribution had a different mean or variance. In this case, we have resolved this issue by applying the test to the standardized residuals rather than the raw residuals.

However, even with this correction, the test does not adjust correctly for the mean or variance of the residuals being *estimated*, which is a critical issue. The book by D'Agostino and Stephens [21] extensively discussed how the distribution of goodness of fit statistics is affected by estimating the mean and variance parameters.

A more comprehensive series of tests is provided by the `EnvStats` package. For example, here are the same two tests within this package:

```
gofTest(lm1$resid,test='sw')
gofTest(lm1$resid,test='ks')
```

with (slightly edited) output:

```
Results of Goodness-of-Fit Test

Test Method:                     Shapiro-Wilk GOF
```

```
Hypothesized Distribution:       Normal
Estimated Parameter(s):          mean = 2.973984e-17
                                 sd   = 7.133667e-01
Estimation Method:               mvue
Data:                            lm1$resid
Sample Size:                     126
Test Statistic:                  W = 0.9913566
Test Statistic Parameter:        n = 126
P-value:                         0.6259324
Alternative Hypothesis:          True cdf does not equal the
                                 Normal Distribution.


Test Method:                     Kolmogorov-Smirnov GOF
Hypothesized Distribution:       Normal
Estimated Parameter(s):          mean = 2.973984e-17
                                 sd   = 7.133667e-01
Estimation Method:               mvue
Data:                            lm1$resid
Sample Size:                     126
Test Statistic:                  ks = 0.04169544
Test Statistic Parameter:        n = 126
P-value:                         0.980819
Alternative Hypothesis:          True cdf does not equal the
                                 Normal Distribution.
Warning message:
In ksGofTest(x = c(-0.952328053800059, 0.349830310085597, -0.388011326028766,  :
  The standard Kolmogorov-Smirnov test is very conservative (Type I error smaller
than assumed; high Type II error) for testing departures from the Normal
distribution when you have to estimate the distribution parameters.
```

The Shapiro-Wilk results are identical to those with `shapiro.test`. The Kolmogorov-Smirnov results are slightly different because this version uses the raw residuals rather than the standardized residuals. Note, however, the `Warning message` at the end: the result does not correctly allow for the estimation of the mean or variance parameters.

Several other test statistics are available within `gofTest`. In particular, the Shapiro-Francia, probability plot correlation coefficient (equivalent to Looney-Gulledge), Cramér-von Mises and Anderson-Darling tests are covered by `test='sf'`, `test='ppcc'`, `test='cvm'` and `test='ad'`, respectively.

Each of the EDF tests is affected by the issue of having to account for estimating the parameters of the distribution, which is usually interpreted as estimating the mean and variance without accounting for the additional parameters involved in a regression equation. To the author's knowledge, there is no systematic theoretical approach to the latter problem. As an alternative, however, we propose a simulation method. Some code for this may be accessed by first entering the following command in a R or R-Studio terminal:

```
source('http://rls.sites.oasis.unc.edu/faculty/rs/source/Data/Rcode-gof.txt')
```

This load a function `gofsim(y,X,intc,msim)` where:

- `y` is the y variable of the regression;
- `X` is the X matrix of the regression;
- `intc` is an indicator for whether an intercept should be included (0 for no, 1 for yes);
- `msim` is the number of simulations.

This function relies on the fact that, for any of the goodness of fit statistics we have considered, the distribution when the null hypothesis of normality is true does not depend on the true values of the regression parameters β, nor, if the residuals are standardized, on the true residual variance σ^2. The first statement follows from (2.56) and the second from the definition of standardized residuals, which is the form of residual used in this routine.

Therefore, to construct a simulated distribution for any of the goodness of fit statistics, it suffices to refit the model to random samples from a standard normal distribution; this is implictly assuming $\beta = \mathbf{0}$ and $\sigma = 1$, but since the distribution is invariant to the true values of β and σ, this will be sufficient.

We should also discuss the value of `msim`, the number of simulations. The examples that follow assume this number is 10,000. Is this adequate? Let's assume that the true p-value of the test is exactly 0.05, which is most commonly taken as the critical value between accepting or rejecting the null hypothesis. We calculate $1.96 \times \sqrt{\frac{0.05 \times 0.95}{10000}} = 0.00427$; thus, there is a better than 95% chance that the simulated p-value will be between 0.045 and 0.055. I would argue that such small differences among p-values does not affect the practical interpretation of the test; in such cases we would know that the result was marginal and further simulation would not give us useful information. Nevertheless, this limitation should be noted.

To conclude this section, I give the results of these tests for both our examples:

```
gofsim(Amherst$Temp,Amherst$Year,1,10000)
gofsim(mta$Charl,mta$MtAiry,1,10000)
```

produce the results

```
[1,] ""         "LG"      "KS"      "CVM"     "AD"
[2,] "Test"     "0.99569" "0.04101" "0.03315" "0.25477"
[3,] "p-value" "0.5582"  "0.8802"  "0.7889"  "0.7306"
..
[1,] ""         "LG"      "KS"      "CVM"     "AD"
[2,] "Test"     "0.99353" "0.07725" "0.07899" "0.461"
[3,] "p-value" "0.4865"  "0.235"   "0.2132"  "0.2538"
```

For example, with the Amherst data, the Looney-Gulledge test has a simulated p-value about 0.56, and the Kolmogorov-Smirnov about 0.88. Because these are simulated p-values, the reader should expect to get slightly different p-values from one run of the routine to another. With neither dataset do any of the four tests point to rejection of the normality hypothesis.