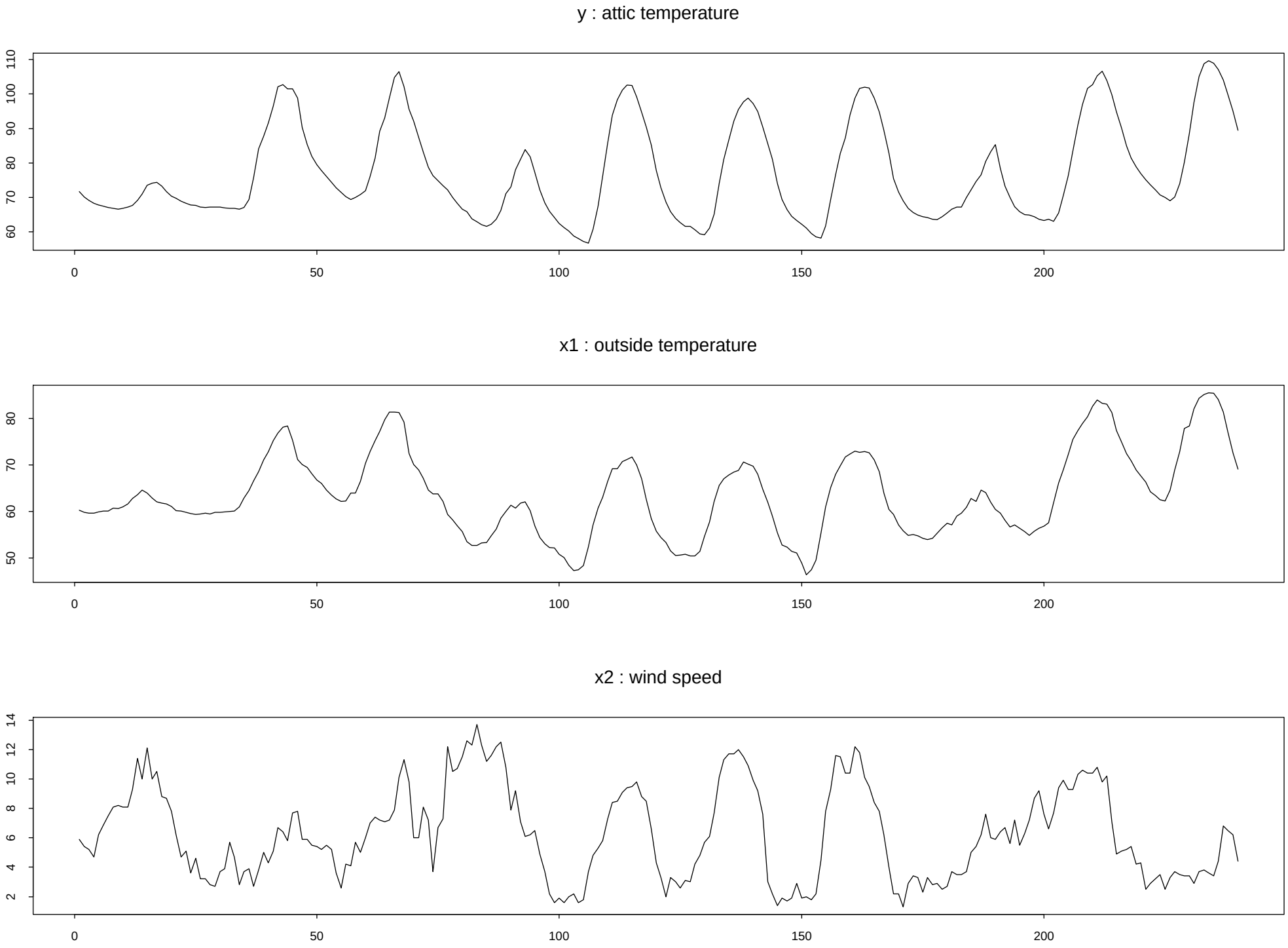
The Attic Data

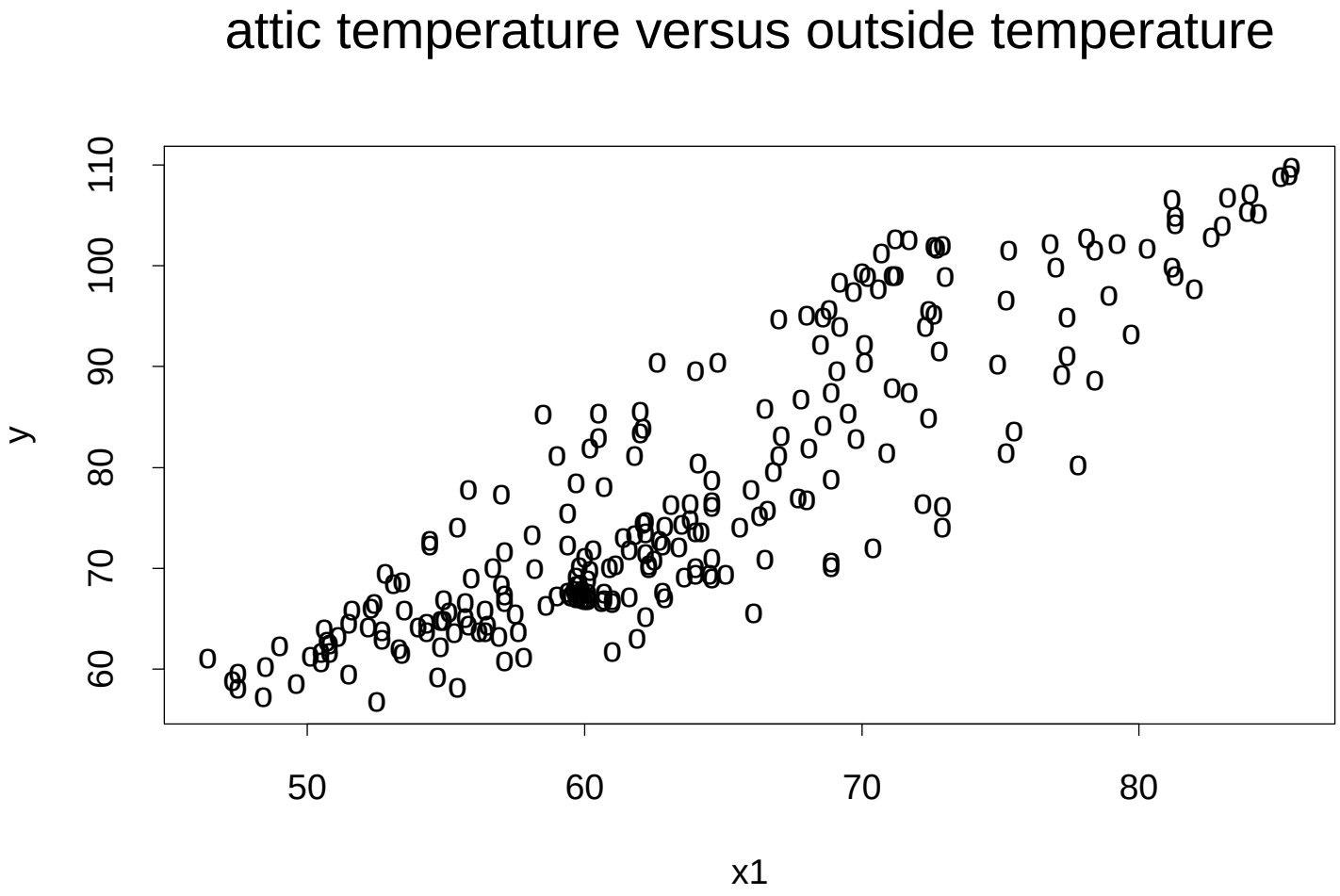
The following time series are hourly measurements (aggregated from 3 twenty-minute readings) taken over 240 consecutive hours from May 29, 1976 to June 7, 1976.

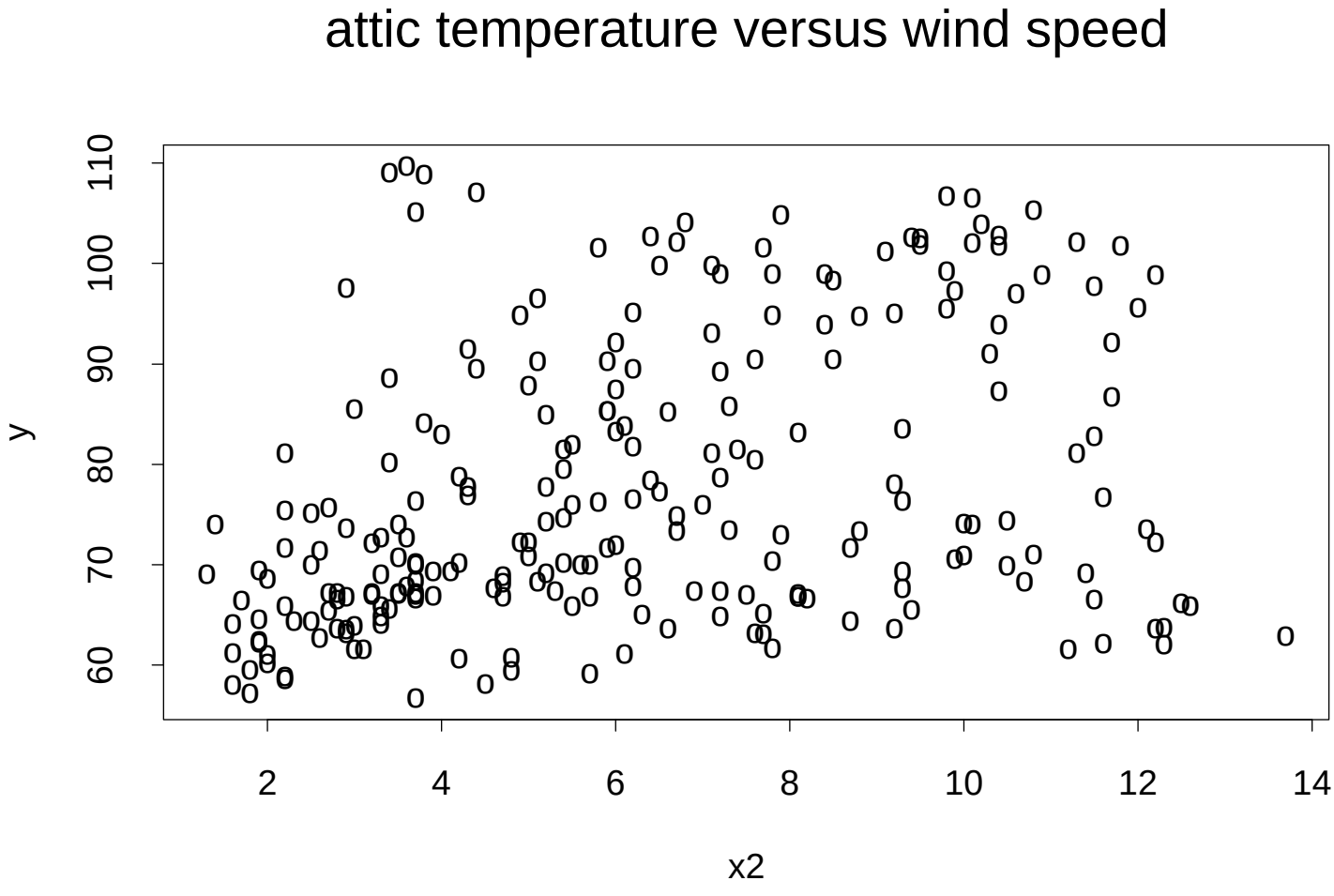
```
   y = attic temperature

   x1 = outside temperature

   x2 = wind speed
```

time	y	x1	x2
1	71.7	60.3	5.9
2	70.1	59.8	5.4
3	69.1	59.7	5.2
4	68.2	59.7	4.7
5	67.8	59.9	6.2
6	67.4	60.1	6.9
7	67.0	60.1	7.5
8	66.8	60.7	8.1
9	66.6	60.6	8.2
10	66.8	61.0	8.1
11	67.1	61.6	8.1
12	67.6	62.8	9.3
13	69.1	63.6	11.4
14	70.9	64.6	10.0
15	73.5	64.0	12.1

```
 16  74.1 62.9 10.0
 17  74.4 62.1 10.5
 18  73.3 61.8  8.8
 19  71.7 61.6  8.7
 20  70.3 61.1  7.8
 21  69.7 60.2  6.2
 22  68.8 60.1  4.7
 23  68.3 59.8  5.1
 24  67.8 59.6  3.6
 25  67.6 59.4  4.6
 26  67.2 59.5  3.2
 27  67.0 59.7  3.2
 28  67.2 59.5  2.8
 29  67.2 59.8  2.7
 30  67.2 59.8  3.7
 31  66.9 59.9  3.9
 32  66.8 60.0  5.7
 33  66.8 60.1  4.7
 34  66.5 61.0  2.8
 35  67.0 62.9  3.7
  .     .    .    .   (observations 36 to 234
  .     .    .    .     are omitted here)
  .     .    .    .
235 109.0 85.4  3.4
236 107.1 84.0  4.4
237 104.1 81.3  6.8
238  99.8 77.0  6.5
239  95.1 72.6  6.2
240  89.5 69.1  4.4
```


y : attic temperature
60
70
80
90
100
110
0
50
100
150
200
x1 : outside temperature
50
60
70
80
0
50
100
150
200
x2 : wind speed
2
4
6
8
10
12
14
0
50
100
150
200


attic temperature versus outside temperature
y
x1
60
70
80
90
100
110
50
60
70
80


attic temperature versus wind speed
y
x2
60
70
80
90
100
110
2
4
6
8
10
12
14

Naive (Wrong) Regression Model

The following is some computer output from fitting the model

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \varepsilon_t\,, \quad t = 1, 2, \ldots, 240,$$

using standard regression (ordinary least squares):

```
-------------------------------------------------
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  -9.4805   3.1025    -3.0558   0.0025
         X1   1.3384   0.0511    26.1843   0.0000
         X2   0.2477   0.1487     1.6652   0.0972

Residual standard error:
   6.685 on 237 degrees of freedom

Multiple R-Squared: 0.7738

Correlation of Coefficients:
   (Intercept)      X1
X1 -0.9489
X2  0.0544     -0.3383
-------------------------------------------------
```

- Ordinary least squares will be reasonable (and optimal) when the "errors" are independent and normally distributed with a mean of zero and a common variance.
- It is clear from the time series plot of the residuals (given later) that the errors are NOT independent. They exhibit considerable serial correlation (also known as autocorrelation).

Consequences of Serial Correlation:

- The coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ will still be unbiased (if serial correlation is the only problem with the model), but are likely to be "inefficient". That is, they will not be as accurate (on average) as estimates produced by methods which take proper account of the serial correlation.
- Most of the rest of the output is just plain wrong! The numbers given could be way off! This remark applies to the standard errors, t-statistics and p-values associated with the coefficient estimates, and also to the reported correlations between the coefficients.

Most regression software includes tests for the presence of serial correlation (such as the well known **Durbin-Watson** test).

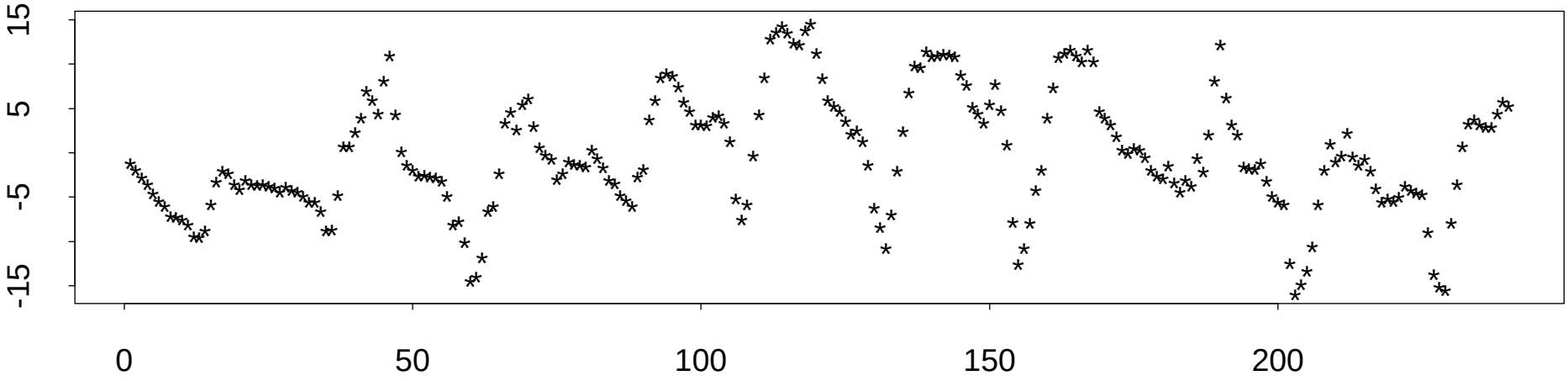
Also note: In situations like this with autocorrelated errors, the inclusion of "lagged" terms in the model may provide a much better explanation of the data and may give superior predictions.

(Using a model with lagged terms may give a much smaller residual standard error.)

Here is an example of a model with "lagged" terms:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{1,t-1} + \beta_3 X_{1,t-2} + \beta_4 X_{2,t} + \beta_5 X_{2,t-1} + \varepsilon_t \, .$$

Residuals from Naive Regression Model



This is what residuals ought to look like!