UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS

Generalised Regression Models

GRM: Example — Multiple regression in R for Barometer data **Semester 1, 2022–2023**

For multiple regression in R use the function `lm`, e.g. for a response vector `y` and explanatory vectors `x1`, `x2`, `x3`, we might have `lm(y ~ x1 + x2 + x3)`.

Example: Barometer data

The data shown below (and given in Barometer.txt) comprise 10 readings of an aneroid barometer (in mm) and corresponding readings of a mercury barometer (mm) and the temperature (°C) and humidity (%) at the time the two barometer readings were taken.

Aneroid barometer	Mercury barometer	Temperature	Humidity
749.0	744.4	10.0	69.1
746.0	741.3	6.2	48.3
756.0	752.7	6.3	50.0
758.9	754.7	5.3	62.7
751.7	747.8	4.8	60.0
757.5	754.0	3.8	31.3
752.4	747.8	17.1	71.4
752.5	748.6	22.2	25.6
752.2	747.7	20.8	30.7
759.5	755.6	21.0	40.2

The mercury barometer measures air pressure by the height of a column of mercury. The aneroid barometer uses the movement of the elastic top of a metal box containing a vacuum, but its scale is graduated in mm of mercury. A relation between their readings (possibly affected by other atmospheric variables) has been suggested based on fitting a regression model of the form

$$\mathrm{E}(Y \mid x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \,, \tag{1}$$

where y, x_1, x_2 and x_3 denote the four variables in the order given above, and β_0, β_1, β_2 and β_3 are unknown parameters. For this model, we may be interested in the following questions.

(a) Are the data consistent with assumptions of linearity, constant variance and Normality?

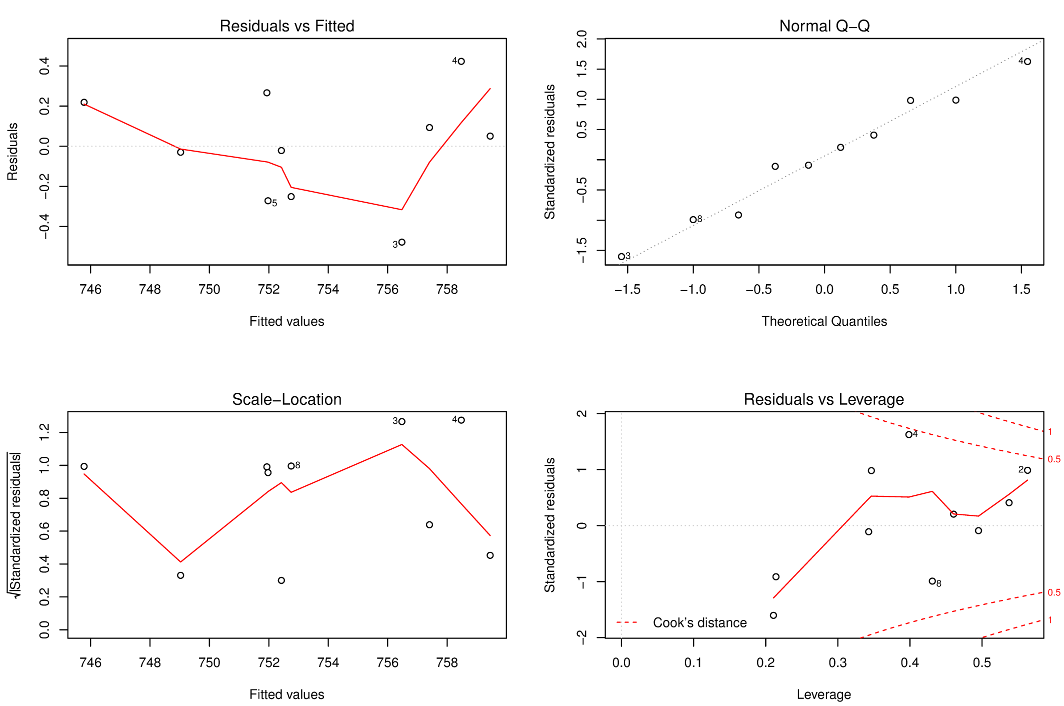
(b) Do the temperature and humidity have any influence on the relationship between the readings of the two barometers?

(c) If not, are the data consistent with the model in which the expected aneroid barometer reading differs from the mercury barometer reading by a constant?

The R output below results from fitting model (1).

```
Barometer.dat <- read.table('Barometer.txt',header=T)
attach(Barometer.dat)
fit1 <- lm(Aneroid.barometer ~ Mercury.barometer + Temperature + Humidity)
summary(fit1)
anova(fit1)
par(mfrow=c(2,2))
plot(fit1)
```

```
Call:
lm(formula = Aneroid.barometer ~ Mercury.barometer + Temperature +
    Humidity)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4777 -0.1956  0.0145  0.1875  0.4229

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       50.978317  18.537485   2.750   0.0333 *
Mercury.barometer  0.936282   0.024585  38.083 2.19e-08 ***
Temperature        0.025454   0.015935   1.597   0.1613
Humidity           0.011996   0.007491   1.601   0.1604
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3354 on 6 degrees of freedom
Multiple R-squared: 0.9961,Adjusted R-squared: 0.9941
F-statistic: 504.6 on 3 and 6 DF,  p-value: 1.344e-07

> Analysis of Variance Table

Response: Aneroid.barometer
                  Df  Sum Sq Mean Sq   F value   Pr(>F)
Mercury.barometer  1 169.911 169.911 1510.1143 1.94e-08 ***
Temperature        1   0.126   0.126    1.1193   0.3308
Humidity           1   0.289   0.289    2.5646   0.1604
Residuals          6   0.675   0.113
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



The regression 'explains' all but $\frac{0.675}{171.001}$ or 0.39% of the total SS about the mean. The residual standard deviation, 0.3354, is the square root of the residual MS, 0.113.

With so few observations, we expect to be able to detect only major violations of the assumptions of the regression model: the Normal probability plot of the standardized residuals and the plot of these residuals against the fitted values provide little evidence that the assumptions are false.

The significance probabilities for tests on the coefficients for temperature and humidity are not small, so there is no evidence that either variable influences the relationship between the two barometers. To test a hypothesis that *neither* variable has an effect, we find the extra SS for omitting both of them.

```
fit2 <- lm(Aneroid.barometer ~ Mercury.barometer)

anova(fit2,fit1)

Analysis of Variance Table

Model 1: Aneroid.barometer ~ Mercury.barometer
Model 2: Aneroid.barometer ~ Mercury.barometer + Temperature + Humidity
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1      8 1.08959
2      6 0.67509  2    0.4145 1.842 0.2378
```

The extra SS and the corresponding MS are 0.415 and 0.415/2 or 0.2075, so an F-statistic for this hypothesis is

$$0.2075/0.113 = 1.84;$$

comparison with $F(2,6)$ shows that the joint influence of these variables is not significant.

We might also test the hypothesis that the true coefficient for the mercury barometer equals 1 by comparing the ratio

$$|0.93628 - 1|/0.0246 = 2.592$$

with $t(6)$: the significance probability (in a two-sided test) is 0.041, so there is slight evidence that this coefficient differs from 1.

The model in which the expected aneroid barometer reading differs from the mercury barometer reading by a constant has $\beta_1 = 1$, $\beta_2 = 0$, $\beta_3 = 0$: this model could be fitted in R using

```
lm(Aneroid.barometer - Mercury.barometer ~ 1)
```

However, the residual SS under this model is simply given by

$$\sum_i \{(y_i - \bar{y}) - (x_{i1} - \bar{x}_1)\}^2 = 2.149,$$

so that the extra SS and the corresponding MS relative to (1) are $2.149 - 0.675$ or 1.474 and 1.474/3 or 0.4913. The F-statistic is 0.4913/0.113 or 4.37; comparison with $F(3,6)$ gives a significance probability of 0.059, showing little evidence against this simplified model.