**Regression Methods: Problems** MATH-408
Anthony Davison

---

**Problem 1** (Interpreting R output)
On fitting the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ to $n = 13$ measures of cement properties, we obtain

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.19363    3.91330  12.315 6.17e-07 ***
x1           1.69589    0.20458   8.290 1.66e-05 ***
x2           0.65691    0.04423  14.851 1.23e-07 ***
x3           0.25002    0.18471   1.354    0.209
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Explain in detail how the columns '`t value`' and '`Pr(>|t|)`' are computed. What do they mean? Comment on the values in the output above.

(b) If $c = (0, 0, 1, -1)^{\mathrm{T}}$, then show that (using an obvious notation)

$$s^2 c^{\mathrm{T}} (X^{\mathrm{T}} X)^{-1} c = \mathrm{SE}\left(\widehat{\beta}_2\right)^2 + \mathrm{SE}\left(\widehat{\beta}_3\right)^2 - 2\,\mathrm{corr}\left(\widehat{\beta}_2, \widehat{\beta}_3\right) \mathrm{SE}\left(\widehat{\beta}_2\right) \mathrm{SE}\left(\widehat{\beta}_3\right).$$

If $\mathrm{corr}(\widehat{\beta}_2, \widehat{\beta}_3) \doteq -0.08911$, give the $p$-value for testing the hypothesis that $\beta_2 = \beta_3$, and say whether it can be rejected at level 5%.

*Reminder*: Recall that, with $c = (0, 0, 1, -1)^{\mathrm{T}}$,

$$s^2 c^{\mathrm{T}} (X^{\mathrm{T}} X)^{-1} c = \left\{\widehat{\mathrm{SE}}\left(\widehat{\beta}_2\right)\right\}^2 + \left\{\widehat{\mathrm{SE}}\left(\widehat{\beta}_3\right)\right\}^2 - 2\,\widehat{\mathrm{corr}}\left(\widehat{\beta}_2, \widehat{\beta}_3\right) \widehat{\mathrm{SE}}\left(\widehat{\beta}_2\right) \widehat{\mathrm{SE}}\left(\widehat{\beta}_3\right).$$

**Problem 2** (Models with factors) In `R`, the general formula for a model is

```
response~expression
```

where the left-hand side, `reponse`, can be missing, the right-hand side, `expression`, is a collection of terms joined by operators, and the full formula is similar to an arithmetic expression. Let

$$y = \begin{pmatrix} 217 \\ 143 \\ 186 \\ 121 \\ 157 \\ 143 \end{pmatrix}, \quad X = \begin{pmatrix} 152 & 1 & 1 \\ 93 & 1 & 2 \\ 127 & 1 & 3 \\ 109 & 2 & 1 \\ 141 & 2 & 2 \\ 136 & 2 & 3 \end{pmatrix},$$

and let `x`, `a`, `b` denote the columns of $X = [x, a, b]$.

By default `R` includes a column of ones as the first column of every design matrix; we call the corresponding parameter $\beta_0$. This column can be suppressed by including `-1` in the model formula.

The *linear predictor* of a model is $\eta = X\beta$, so $\eta_j = x_j^{\mathrm{T}} \beta$ corresponds to the $j$th observation.

(a) A *factor* represents a categorical observation (command `as.factor()` in `R`). For instance, if `a` is a factor, then `y~a` gives

$$\eta_j = \beta_0 + \alpha_1, \quad j = 1, 2, 3, \qquad \eta_j = \beta_0 + \alpha_2, \quad j = 4, 5, 6,$$

where $\beta_0$, $\alpha_1$ et $\alpha_2$ are parameters. Alternatively we can use indicator functions and write

$$\eta_j = \beta_0 + \alpha_1 I_{(a_j=1)} + \alpha_2 I_{(a_j=2)}, \tag{1}$$

where $I_E = 1$ if the condition $E$ is true, and 0 otherwise. The values "1" and "2" in a factor `a` do not represent the numbers 1 and 2, but categories, groups, classes or levels. For instance, `a` could represent "1" = "regular food regime", and "2" = "food regime with growth inhibitors".

If `a` and `b` are factors,

(i) give the design matrix and the vector of parameters for the model (1).

(ii) This design matrix is not full rank. What consequence has this for estimation?

(iii) Delete the column corresponding to $\alpha_1$ to make the matrix full rank. What is now the interpretation of $\beta_0$ and $\alpha_2$?

(iv) When the model includes a constant $\beta_0$, `R` automatically suppresses the first level of every factor. Give the design matrices for the following formulae

`y~a`, `y~a+b`, `y~x+a-1`, `y~b+x-1`.

Note that the + (and `-`) in these expressions indicates addition (and removal) of the vector subspaces spanned by the terms, not to 'ordinary' addition.

(b) If `a` and `b` are factors, an *interaction* component is represented by `a:x` or `a:b`. For instance, `y~a:x` gives

$$\eta_j = \beta_0 + \alpha_1 x_j + \varepsilon_j, \quad j = 1, 2, 3, \qquad \eta_j = \beta_0 + \alpha_2 x_j + \varepsilon_j, \quad j = 4, 5, 6,$$

which can also be written

$$\eta_j = \beta_0 + \alpha_1 I_{(a_j=1)} x_j + \alpha_2 I_{(a_j=2)} x_j;$$

this gives different slopes for the groups "1" and "2", but a common intercept.

Similarly, the expression `y~a:b` represents the model

$$\eta_j = \beta_0 + \alpha_j, \quad j = 1, \ldots, 6,$$

which can also be written

$$\eta_j = \beta_0 + \sum_{r=1}^{2} \sum_{s=1}^{3} \gamma_{r,s} I_{(a_j=s)} I_{(b_j=r)},$$

i.e., a model with different intercepts for every combination of levels of `a` and `b`.

Give the design matrices for the formulae

`y~a:x`, `y~a:b` , `y~a+b:x`, `y~a+a:b:x`,

and say which of them have linearly independent columns.

*Hint:* You can check your answers using the `R` commands:

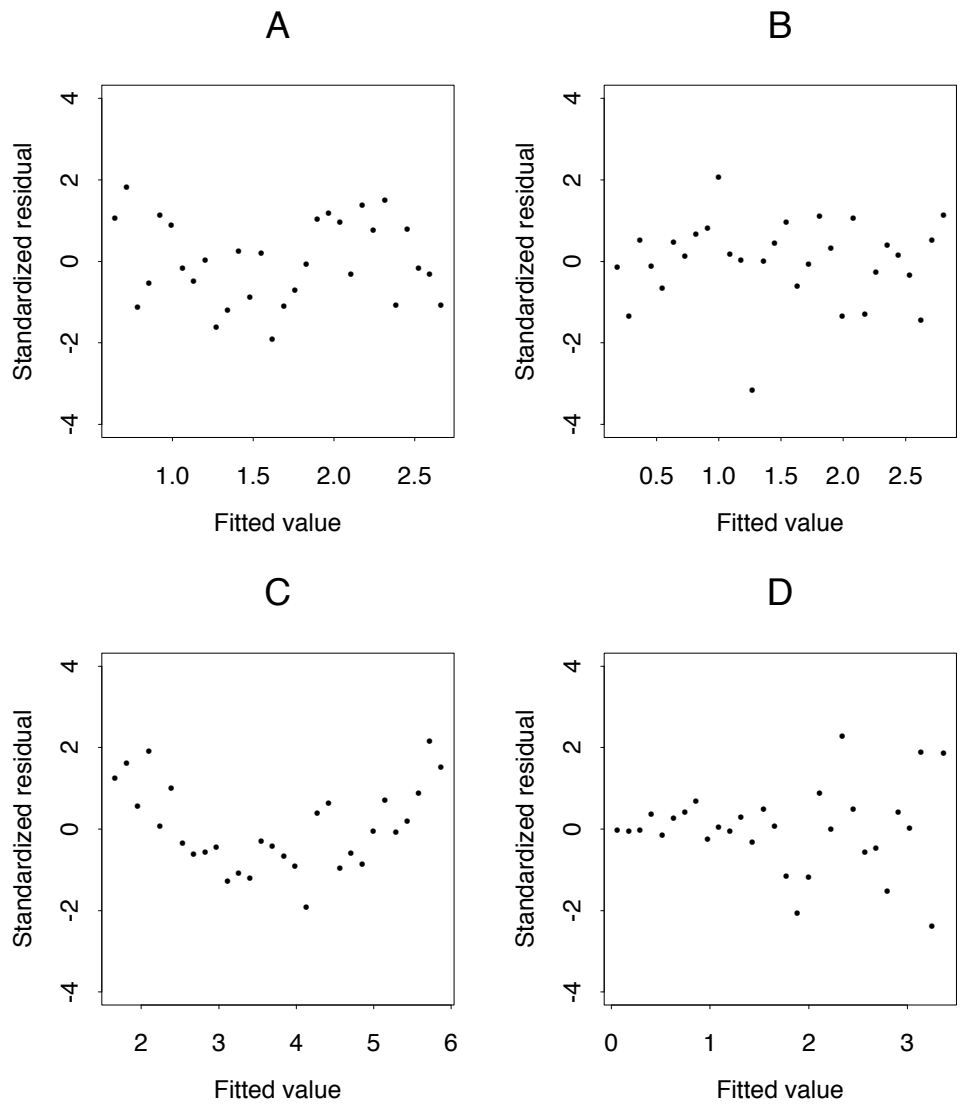


Figure 1: Standardized residuals for four Gaussian linear models.

```
y <- c(217,143,186,121,157,143)
X <- matrix(c(152,93,127,109,141,136,1,1,1,2,2,2,1,2,3,1,2,3),6,3)
df <- data.frame(y = y, x = X[,1], a = as.factor(X[,2]), b = as.factor(X[,3]))
model.matrix(reponse~expression, data = df)
```

**Problem 3** (Graphical diagnostics)

(a) Figure 1 shows standardized residuals for four different datasets. Discuss each fit and explain briefly how any problem might be fixed.

(b) Figure 2 shows four Gaussian Q-Q plots, for data with (i) heavier tails than the Gaussian; (ii) lighter tails than the Gaussian; (iii) positive skewness; and (iv) negative skewness. Match these with the panels of Figure 2, explaining your reasoning.

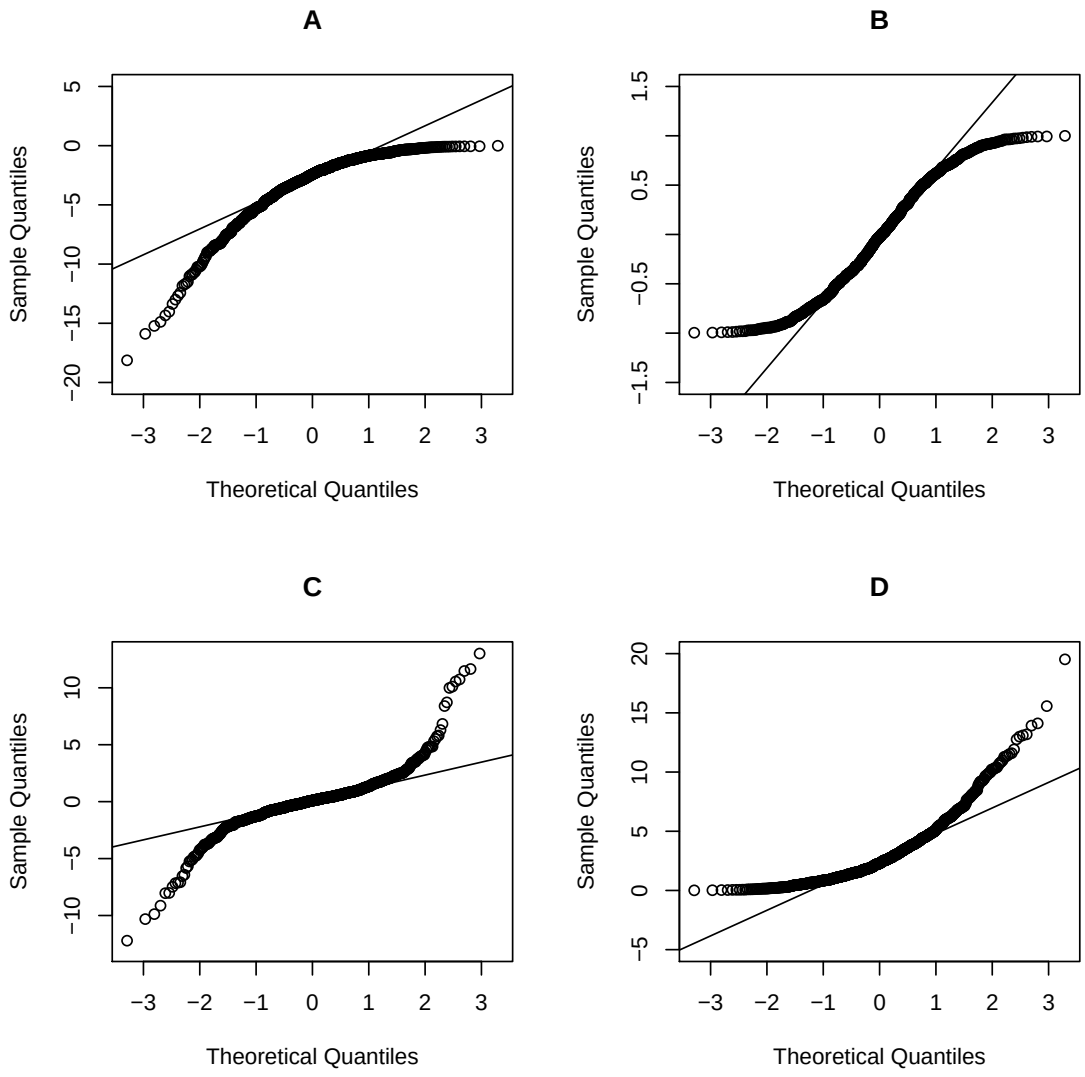


Figure 2: Four Gaussian Q-Q plots.