Linear model with categorical predictors

Suparna

2025-10-17

One factor four levels: An example

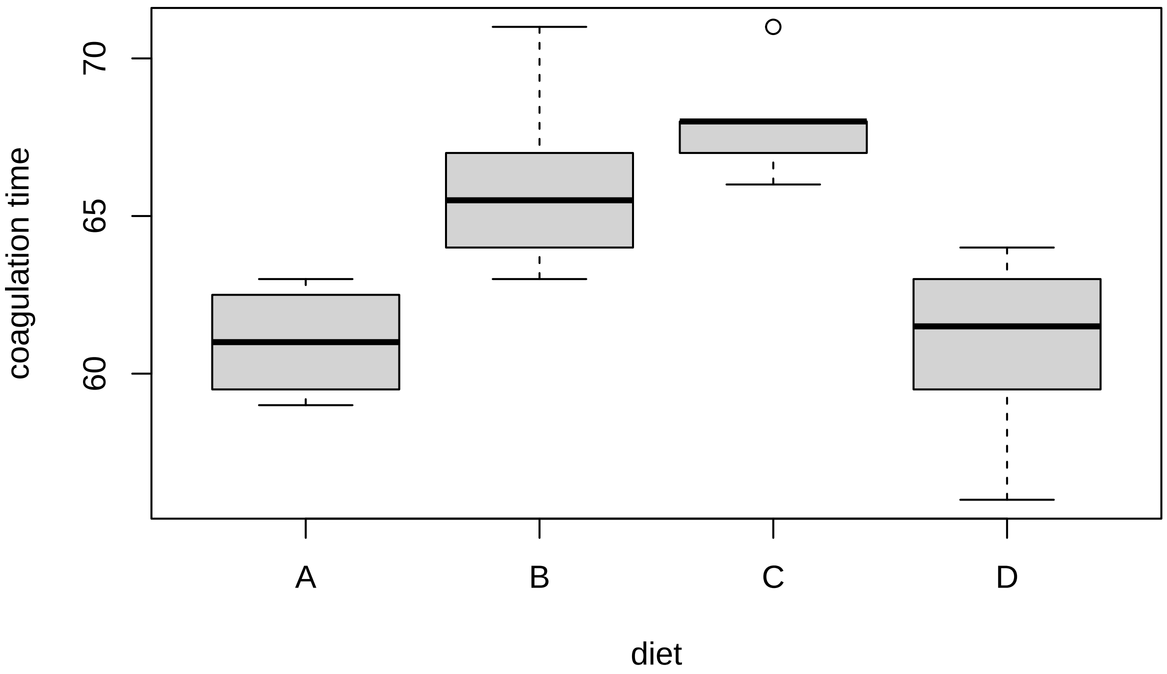
Twenty-four animals were randomly assigned to four different diets and blood samples were taken in a random order. The blood coagulation time was measured in seconds. These data come from Box et al. (1978):

```
library(faraway)
data(coagulation, package="faraway")
coagulation
```

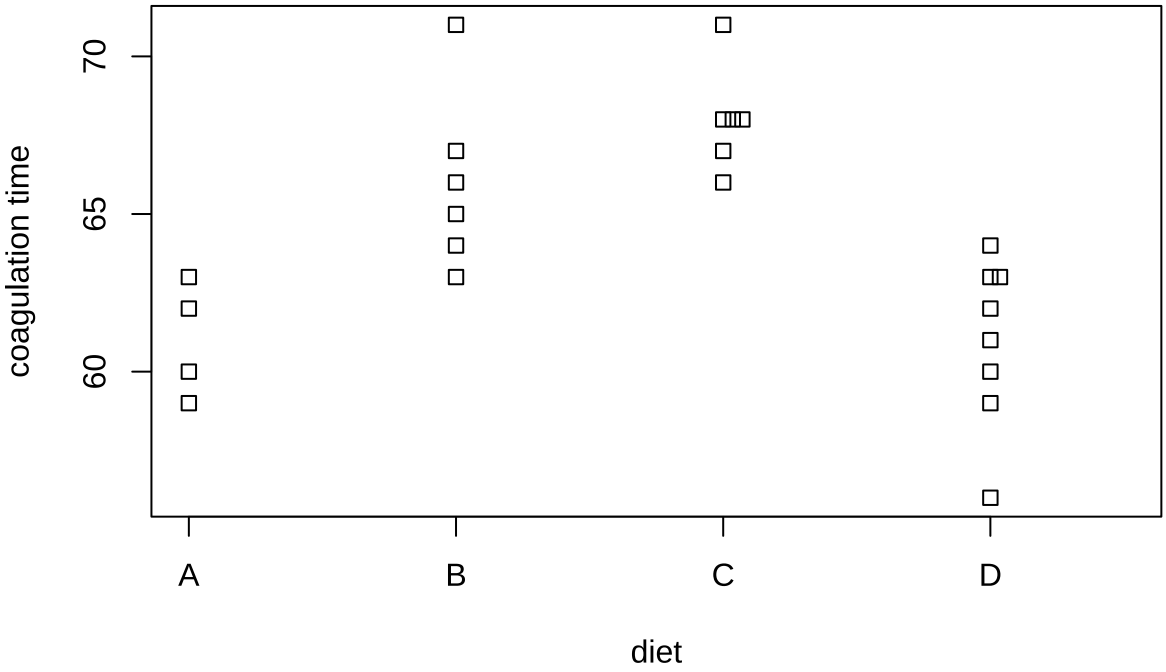
```
##    coag diet
## 1    62    A
## 2    60    A
## 3    63    A
## 4    59    A
## 5    63    B
## 6    67    B
## 7    71    B
## 8    64    B
## 9    65    B
## 10   66    B
## 11   68    C
## 12   66    C
## 13   71    C
## 14   67    C
## 15   68    C
## 16   68    C
## 17   56    D
## 18   62    D
## 19   60    D
## 20   61    D
## 21   63    D
## 22   64    D
## 23   63    D
## 24   59    D
```

Some preliminary graphical analysis is essential before fitting. So first we are gonna check the side-by-side boxplot. Also, we are gonna check the stripchart, since a stripchart can be better for smaller datasets:

```r
plot(coag ~ diet, coagulation,ylab="coagulation time")
```



```r
stripchart(coag ~ diet, coagulation, vertical=TRUE, method="stack",
xlab="diet",ylab="coagulation time")
```



The boxplot shows how the four levels vary but there is something odd about the display of diet C because the median and upper quartile are the same. The stripchart shows the ties in the data in diets C and D.

We must check for equality of variance in the groups, which seems satisfied in this example. We are looking for evidence of skewness showing a lack of normality. This might suggest a transformation of the response. There is no such concern in this example. Finally, we should look for outliers —there are none to be seen in this example.

Now let's fit the model using the default treatment coding:

```
lmod <- lm(coag ~ diet, coagulation)
sumary(lmod)
```

```
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 6.1000e+01 1.1832e+00 51.5544 < 2.2e-16
## dietB       5.0000e+00 1.5275e+00  3.2733 0.0038025
## dietC       7.0000e+00 1.5275e+00  4.5826 0.0001805
## dietD       2.7195e-15 1.4491e+00  0.0000 1.0000000
##
## n = 24, p = 4, Residual SE = 2.36643, R-Squared = 0.67
```

Rounding error results in a rather unfortunate formatting of the results. The coefficients can be more cleanly seen as:

```
round(coef(lmod),1)
```

```
## (Intercept)       dietB       dietC       dietD
##          61           5           7           0
```

Group A is the reference level and has a mean of 61, groups B, C and D are 5, 7 and 0 seconds larger, respectively, on average. Examine the design matrix to understand the coding:

```r
model.matrix(lmod)
```

```
##    (Intercept) dietB dietC dietD
## 1            1     0     0     0
## 2            1     0     0     0
## 3            1     0     0     0
## 4            1     0     0     0
## 5            1     1     0     0
## 6            1     1     0     0
## 7            1     1     0     0
## 8            1     1     0     0
## 9            1     1     0     0
## 10           1     1     0     0
## 11           1     0     1     0
## 12           1     0     1     0
## 13           1     0     1     0
## 14           1     0     1     0
## 15           1     0     1     0
## 16           1     0     1     0
## 17           1     0     0     1
## 18           1     0     0     1
## 19           1     0     0     1
## 20           1     0     0     1
## 21           1     0     0     1
## 22           1     0     0     1
## 23           1     0     0     1
## 24           1     0     0     1
## attr(,"assign")
## [1] 0 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$diet
## [1] "contr.treatment"
```

To specifically answer the question of whether there is a significant difference between any of the levels we need to look at the anova table

```r
anova(lmod)
```

```
## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       3    228    76.0  13.571 4.658e-05 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that there is indeed a difference in the levels although this test does not tell us which levels are different from others.

We can fit the model without an intercept term as in:

```r
lmodi <- lm(coag ~ diet -1, coagulation)
sumary(lmodi)
```

```
##       Estimate Std. Error t value  Pr(>|t|)
## dietA 61.00000    1.18322  51.554 < 2.2e-16
## dietB 66.00000    0.96609  68.317 < 2.2e-16
## dietC 68.00000    0.96609  70.387 < 2.2e-16
## dietD 61.00000    0.83666  72.909 < 2.2e-16
##
## n = 24, p = 4, Residual SE = 2.36643, R-Squared = 1
```

We can directly read the level means. The R2 is not correctly calculated because of the absence of an intercept. To generate the usual test that the means of the levels are equal, we would need to fit the null model and compare using an F-test:

```
lmnull <- lm(coag ~ 1, coagulation)
summary(lmnull)
```

```
##
## Call:
## lm(formula = coag ~ 1, data = coagulation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.00  -2.25  -0.50   3.00   7.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.0000     0.7848   81.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.845 on 23 degrees of freedom
```

```
anova(lmnull,lmodi)
```

```
## Analysis of Variance Table
##
## Model 1: coag ~ 1
## Model 2: coag ~ diet - 1
##   Res.Df RSS Df Sum of Sq      F    Pr(>F)
## 1     23 340
## 2     20 112  3       228 13.571 4.658e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We get the same F-statistic and p-value as in the first coding.

We can also use a sum coding:

```
options(contrasts=c("contr.sum","contr.poly"))
lmods <- lm(coag ~ diet , coagulation)
sumary(lmods)
```

```
##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 64.00000    0.49791 128.5367 < 2.2e-16
## diet1       -3.00000    0.97361  -3.0813 0.0058890
## diet2        2.00000    0.84533   2.3659 0.0281950
## diet3        4.00000    0.84533   4.7319 0.0001276
##
## n = 24, p = 4, Residual SE = 2.36643, R-Squared = 0.67
```

So the estimated overall mean response is 64 while the estimated mean response for A is three less than the overall mean, that is, 61. Similarly, the means for B and C are 66 and 68, respectively. Since we are using the sum constraint, we compute $\hat{\alpha}_D = -(-3 + 2 + 4)$ so the mean for D is 64-3=61. Notice that $\hat{\sigma}$ and R^2 are the same as before.

So we can use any of these three methods and obtain essentially the same results. Dropping the intercept is least convenient since an extra step is needed to generate the F-test. Furthermore, the approach would not extend well to experiments with more than one factor, as additional constraints would be needed. The other two methods can be used according to taste. The treatment coding is most appropriate when the reference level is set to a possible control group.