2.8 Extension

Non-additive associations

The additive assumption means that *the effect of changes in a predictor X_j on the response Y is independent of the values of other predictors.*

In many situations, there is a **synergy** effect, i.e. increasing the level of one covariate may interact with the level of another. This is called **interaction** in statistics.

Example

Take

$$\mathrm{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

This means that

$$\mathrm{Y} = \beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2 + \varepsilon = \beta_0 + \tilde{\beta}_1 x_1 + \beta_2 x_2 + \varepsilon$$

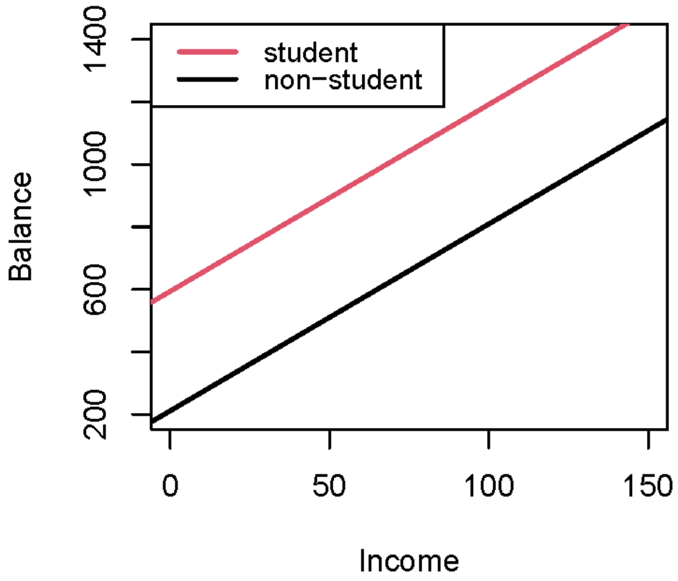
where $\tilde{\beta}_1 = \beta_1 + \beta_3 x_2$, i.e. $\tilde{\beta}_1$ changes with x_2 and the effect of x_1 on Y is no longer constant.

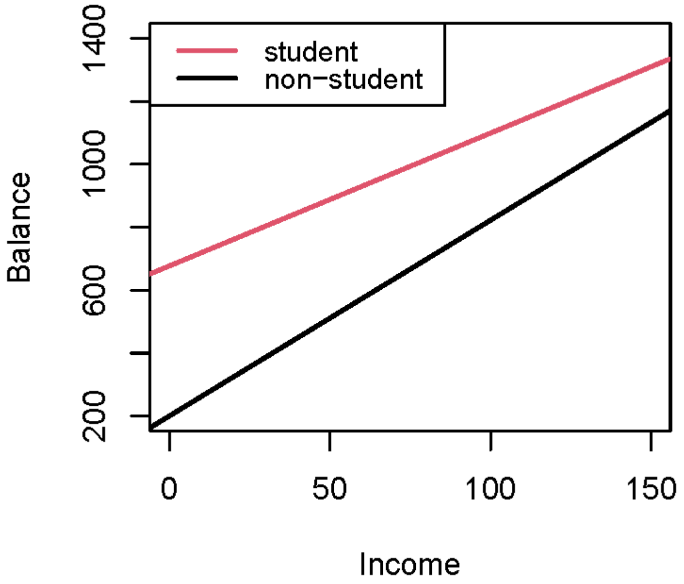
Remark:

- The **hierarchical principle** states that if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.
- The concept of interactions applies to qualitative variables, to quantitative variables or to a combination of both.

Example

Using the `Credit` data (`ISLR` package), the least squares lines are shown for prediction of `Balance` from `Income` for students and non-students. Top model does not include interactions, bottom model does.






```
library(ISLR)

data("Credit")

lm <- lm(Balance~Income+Student, data=Credit)

plot(0,0, xlim=c(0,150), ylim=c(200,1400), xlab="Income", ylab="Balance")
abline(a=lm$coefficients[1]+lm$coefficients[3], b=lm$coefficients[2], col=2,
       lwd=2)
abline(a=lm$coefficients[1], b=lm$coefficients[2], col=1, lwd=2)
legend("topleft", legend=c("student", "non-student"), col=c(2,1), lwd=2,
       cex=0.9)

lm2 <- lm(Balance~Income*Student, data=Credit)

plot(0,0, xlim=c(0,150), ylim=c(200,1400), xlab="Income", ylab="Balance")
abline(a=lm2$coefficients[1]+lm2$coefficients[3],
       b=lm2$coefficients[2]+lm2$coefficients[4], col=2, lwd=2)
abline(a=lm2$coefficients[1], b=lm2$coefficients[2], col=1, lwd=2)
legend("topleft", legend=c("student", "non-student"), col=c(2,1), lwd=2,
       cex=0.9)
```

Non-linear associations

So far, we have considered only linear associations between $\mathbf{X}$ and Y, but in other situations the relationship can be non-linear.

A non-linear association can be suggested by looking at the residuals.

A popular model is a **U-shaped association**, that can be modelled by a quadratic association

$$\mathbb{E}(\mathrm{Y}_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

This is still a **linear regression** because the regression equation is still a linear combination of the explanatory variables X and X^2.

Tips:

- **centre** the explanatory variables
- **scale** the explanatory variables

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\mathrm{sd}(x)}$$

Scaling has several advantages

- numerical accuracy of matrix manipulation is improved, in particular in presence of large values of the covariate
- the intercept β_0 relates the average of y to the average of x, instead of the average of y with $x = 0$ (which is sometimes an impossible value)
- the slope represents a one standard deviation change which is potentially more meaningful than a one unit change (which can be very small or very large)

Fractional polynomials

The quadratic function is symmetric, however there may be situations where *the rate of increase is faster than the rate of decrease*.

A range of functions can be investigated through **fractional polynomials**

$$\mathbb{E}(\mathrm{Y}_i) = \beta_0 + \beta_1 x_i^p \qquad p \neq 0$$

In this case, you can test several models ($p = 1$ is *linear*, $p = 2$ is *quadratic*, $p = -2$ is *reciprocal quadratic*) and investigate the best fit. If $p = 0$, we use $\log(x_i)$.

Since the curves of association can be modified by both the function and the slope parameter, a large number of potential non-linear association can be investigated.