**Subgroup analysis**, with a factor predictor and a continuous covariate, is a form of interaction modelling; the factor predictor *interacts* with the covariate to modify the slope across the subgroups, for example.

We can describe the models using the notation previously introduced for ANOVA; consider the single binary factor predictor and single covariate case;

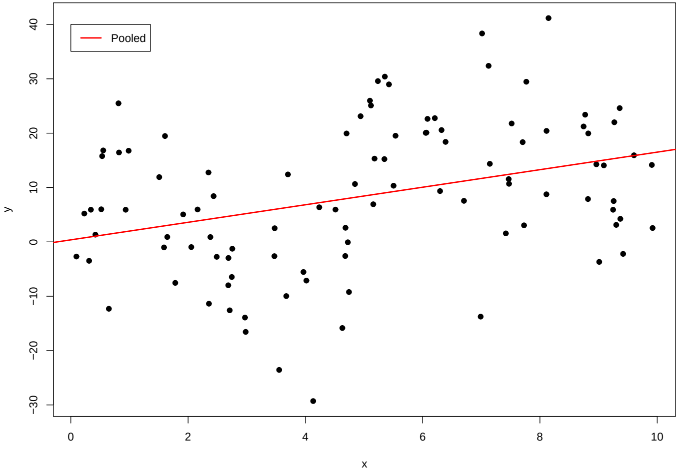
| | | |
|---|---|---|
| MODEL 0 | Single horizontal straight line | $1$ |
| MODEL 1 | Two parallel horizontal straight lines | $X_2$ |
| MODEL 2 | Single straight line, non-zero slope | $X_1$ |
| MODEL 3 | Two parallel straight lines, non-zero slope | $X_1 + X_2$ |
| MODEL 4 | Two non-parallel straight lines | $X_1 + X_2 + X_1.X_2$ |

Note: Always be on the lookout for *lurking* subgroups (subgroups determined by the levels of an unnoticed factor predictor)

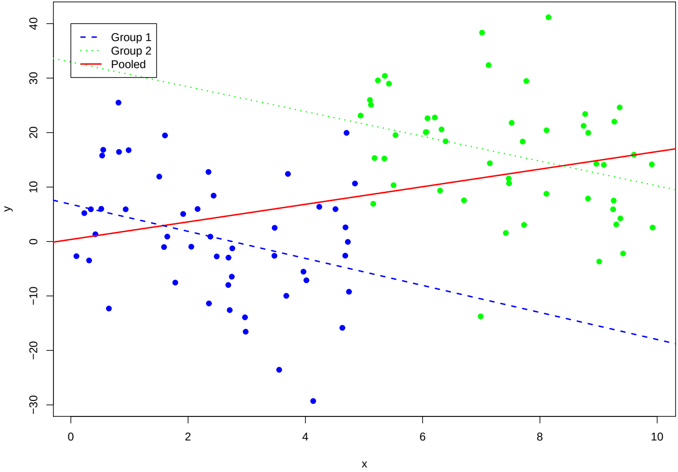
Inferences can change radically when the lurking factor is included in the model

- positive association can be converted into negative association with the continuous covariate.

For example, for factor predictor $X_2$ taking two levels and continuous covariate $X_1$. When the pooled data are examined, a **positive association** between $Y$ and $X_1$ is revealed.



When the pooled data are separated into subgroups, a **negative association** between $Y$ and $X_1$ in each subgroup is revealed.




$X_2 = 0$ in blue, $X_2 = 1$ in green.

i.e. increasing $X_1$ decreases response in subgroup 1, and decreases response in subgroup 2, but appears to increase response overall.

This is known as **Simpson's Paradox in Regression**. It illustrates that pooling data over subgroups must be carried out with care !

- you must fit the factor predictor in the model if you suspect subgroup differences exist.

In the example, the problem arises due to **dependence** between $X_1$ and $X_2$; all the group with $X_2 = 0$ have **low** values of $X_1$, whereas all the group with $X_2 = 1$ have **high** values of $X_1$

Dependence between covariates and factor predictors makes model fitting and results interpretation complicated.

Recap: we can build general models

$$y_i = \beta_0 + \sum_{j=1}^{k} x_{ij} + \epsilon_i$$

to explain the variation of $y$ in terms of covariates and factor predictors $x_1, \ldots, x_k$.

- Simple Linear Regression
- Polynomial Regression
- Multiple Regression
- Factor Predictor Regression
- Interaction Models

We can fit each of these models easily using least-squares to obtain

- estimates $\underset{\sim}{\widehat{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k)^\mathsf{T}$
- standard errors
- goodness of fit measures $R^2$ and Adjusted $R^2$
- residuals for model checking
- predictions

# Interpreting $\widehat{\beta}_j$



$\widehat{\beta}_j$ can be interpreted as the amount of increase in response $y$ when $x_j$ increases by one unit when the other predictors

$$x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$$

are held fixed.

We can test the hypothesis

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_0 &: \beta_j \neq 0 \end{aligned}$$

using the usual hypothesis testing approach.

Test statistic:

$$t_j = \frac{\widehat{\beta}_j}{s_{\widehat{\beta}_j}} = \frac{\text{ESTIMATE}}{\text{STANDARD ERROR}}$$

If $H_0$ is **true**,

$$t_j \sim Student(n - k - 1)$$

as we are estimating $k + 1$ parameters overall.

Note: In multiple regression, when testing each of

$$\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$$

we should strictly use a **multiple testing correction** (as in post-hoc tests in ANOVA)