Level M Regression Models Lecture 13

We have seen how to specify a regression model with factors and interactions. We were interested in whether or not we see a signficantly different relationship between the response and explanatory variable for each group (level of a factor).

R code

Please find some R code used to fit and plot some models. You can find most data sets on Moodle. Please download data and try some of the R code as you read through these notes.

Regression models with factors and interactions

We have already discussed three natural models expressed as:

1. A collection of different regression lines (a model which includes an interaction with a factor),
2. A collection of parallel regression lines,
3. A single regression line (with no differences among the groups).

using the notation:
y_{ij} : response observation j in group i
x_{ij} : explanatory variable observation j in group i
n_i : sample size in group i
p : number of groups
n : $\sum_{i=1}^{p} n_i$, total sample size.
The most general model (1), could be formulated as

$$E(Y_{ij}) = \alpha_i + \beta_i x_{ij}$$

as group i has its own slope and intercept.

In previous weeks we have found that the formulation in terms of $(x_i - \bar{x})$ led to simpler algebra. This is also true in the present case and we will therefore formulate the model as

$$E(Y_{ij}) = \alpha_i + \beta_i(x_{ij} - \bar{x}_{i.})$$

or

$$Y_{ij} = \alpha_i + \beta_i(x_{ij} - \bar{x}_{i.}) + \epsilon_{ij}$$

where $\bar{x}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$, the mean of the explanatory variable, x_{ij} for group i.
The models of interest can be expressed as:

1. different lines: $E(Y_{ij}) = \alpha_i + \beta_i(x_{ij} - \bar{x}_{i.})$
2. parallel lines: $E(Y_{ij}) = \alpha_i + \beta(x_{ij} - \bar{x}_{i.})$
3. single line: $E(Y_{ij}) = \alpha + \beta(x_{ij} - \bar{x}_{..})$, where $\bar{x}_{..}$ is the mean of all x_{ij}.

Note that the "single line" model is written in terms of $(x_{ij} - \bar{x}..)$ so that there are no differences among the groups.

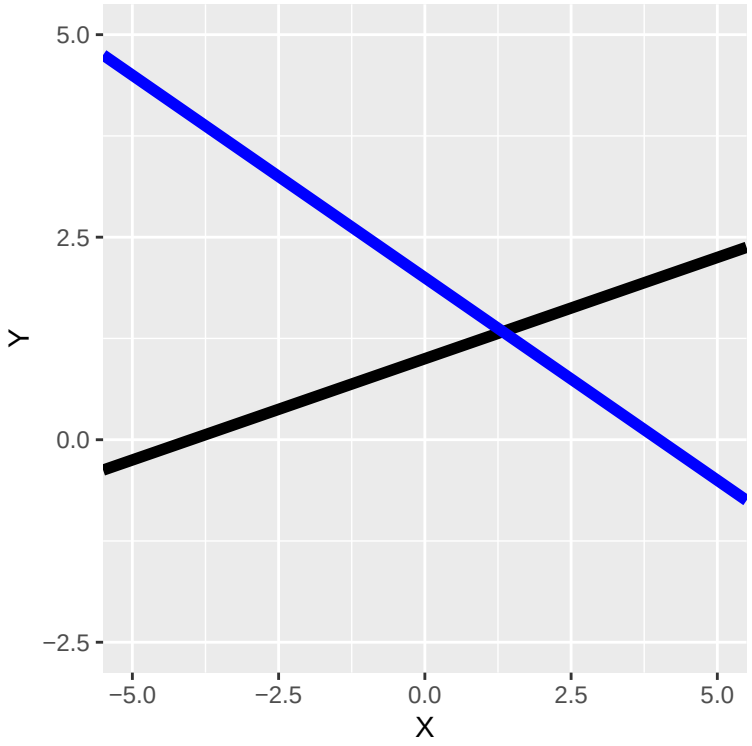
We will now consider how to select an 'appropriate' model given such data. We will begin with the most general model with different regression lines for each group. Let's assume that $i = 2$ such that we have two groups.

Different lines

In order to compare our three models of interest we can firstly assume the more general model 1 and examine whether it is reasonable to adopt the model which has parallel lines, model 2. We can do this by constructing a 95% confidence interval (C.I.) for $(\beta_1 - \beta_2)$, that is the difference between the two slope parameters.

Recall the only difference between model 1 and model 2 is that model 1 contains two slope paramters β_1 and β_2 and model 2 has one slope parameter β. In other words, model 2 assumes $\beta = \beta_1 = \beta_2$.

For example, in the plot below we have two lines (blue and black). It is clear here that the black line has a positive slope parameter and the blue line has a negative slope parameter. This is a clear example of two lines with different slope and intercept terms.



Model 1: Different slope and intercept terms

The most general way to tackle fitting these lines is to formulate each in matrix form and to use the results we have already derived.

$$
\begin{aligned}
E(Y_{11}) &= \alpha_1 + \beta_1(x_{11} - \bar{x}_{1.}) \\
&\vdots \\
E(Y_{1n_1}) &= \alpha_1 + \beta_1(x_{1n_1} - \bar{x}_{1.}) \\
E(Y_{21}) &= \alpha_2 + \beta_2(x_{21} - \bar{x}_{2.}) \\
&\vdots \\
E(Y_{pn_p}) &= \alpha_p + \beta_p(x_{pn_p} - \bar{x}_{p.})
\end{aligned}
$$

i.e. $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ where

$$
\begin{aligned}
\mathbf{Y} &= \begin{pmatrix} y_{11} \\ . \\ . \\ y_{1n_1} \\ y_{21} \\ . \\ . \\ y_{2n_2} \end{pmatrix} \\
\mathbf{X} &= \begin{pmatrix} 1 & (x_{11} - \bar{x}_{1.}) & 0 & 0 \\ . & . & . & . \\ . & . & . & . \\ 1 & (x_{1n_1} - \bar{x}_{1.}) & 0 & 0 \\ 0 & 0 & 1 & (x_{21} - \bar{x}_{2.}) \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & 1 & (x_{2n_2} - \bar{x}_{2.}) \end{pmatrix} \\
\boldsymbol{\beta} &= \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix}
\end{aligned}
$$

The least-squares estimate for $\boldsymbol{\beta}$ is therefore given by

$$
\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}
$$

$$
\begin{aligned}
\mathbf{X}^T\mathbf{X} &= \begin{pmatrix} n_1 & \sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{1.}) & 0 & 0 \\ \sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{1.}) & \sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{1.})^2 & 0 & 0 \\ 0 & 0 & n_2 & \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_{2.}) \\ 0 & 0 & \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_{2.}) & \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_{2.})^2 \end{pmatrix} \\
&= \begin{pmatrix} n_1 & 0 & 0 & 0 \\ 0 & S_{x_1x_1} & 0 & 0 \\ 0 & 0 & n_2 & 0 \\ 0 & 0 & 0 & S_{x_2x_2} \end{pmatrix}
\end{aligned}
$$

since $\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.}) = 0$ and $S_{x_1x_1} = \sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{1.})^2$.

$$\begin{aligned}
(\mathbf{X}^T\mathbf{X})^{-1} &= \begin{pmatrix} \frac{1}{n_1} & 0 & 0 & 0 \\ 0 & \frac{1}{S_{x_1x_1}} & 0 & 0 \\ 0 & 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & 0 & \frac{1}{S_{x_2x_2}} \end{pmatrix} \\
\mathbf{X}^T\mathbf{Y} &= \begin{pmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{1.})y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \\ \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_{2.})y_{2j} \end{pmatrix} \\
&= \begin{pmatrix} n_1\bar{y}_{1.} \\ S_{x_1y_1} \\ n_2\bar{y}_{2.} \\ S_{x_2y_2} \end{pmatrix}
\end{aligned}$$

Therefore,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \bar{y}_{1.} \\ \frac{S_{x_1y_1}}{S_{x_1x_1}} \\ \bar{y}_{2.} \\ \frac{S_{x_2y_2}}{S_{x_2x_2}} \end{pmatrix}$$

Also

$$\begin{aligned}
RSS &= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij}^2 - n_1\bar{y}_{1.}^2 - \frac{(S_{x_1y_1})^2}{S_{x_1x_1}} - n_2\bar{y}_{2.}^2 - \frac{(S_{x_2y_2})^2}{S_{x_2x_2}} \\
&= S_{y_1y_1} - \frac{(S_{x_1y_1})^2}{S_{x_1x_1}} + S_{y_2y_2} - \frac{(S_{x_2y_2})^2}{S_{x_2x_2}} \\
&= RSS_1 + RSS_2
\end{aligned}$$

where RSS_i is the residual sum-of-squares from a simple linear regression fitted to group i.

We therefore verify that the parameter estimates are identical to those obtained by fitting a regression line to each group separately.

95% Confidence interval for slope parameters for two regression lines

Model : $E(Y_{ij}) = \alpha_i + \beta_i(x_{ij} - \bar{x}_{i.})$

Calculate the 95% confidence interval (CI) for $(\beta_1 - \beta_2)$.

This quantity of interest $(\beta_1 - \beta_2)$ can be written as $\mathbf{b}^T\boldsymbol{\beta}$ where

$$\mathbf{b}^T = \begin{pmatrix} 0 & 1 & 0 & -1 \end{pmatrix}.$$

The standard formula now applies.

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t(n-p;0.975)\sqrt{\frac{RSS}{n-p}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

and

$$\begin{aligned}
\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b} &= \begin{pmatrix} 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{n_1} & 0 & 0 & 0 \\ 0 & \frac{1}{S_{x_1x_1}} & 0 & 0 \\ 0 & 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & 0 & \frac{1}{S_{x_2x_2}} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \end{pmatrix} \\
&= \begin{pmatrix} 0 & \frac{1}{S_{x_1x_1}} & 0 & -\frac{1}{S_{x_2x_2}} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \end{pmatrix} \\
&= \frac{1}{S_{x_1x_1}} + \frac{1}{S_{x_2x_2}}
\end{aligned}$$

i.e. a 95% C.I. for $\beta_1 - \beta_2$ is

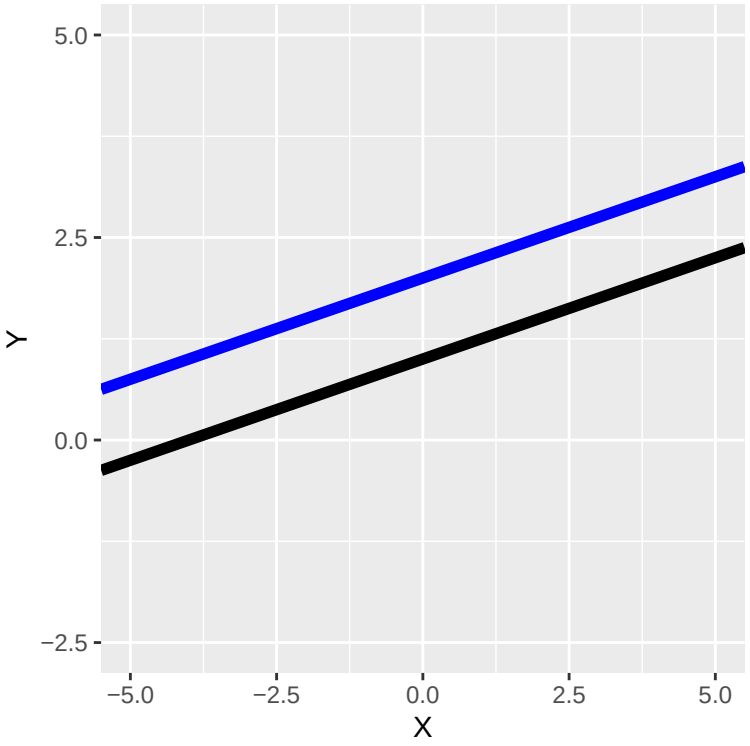
$$\hat{\beta}_1 - \hat{\beta}_2 \pm t(n_1+n_2-4;0.975)\sqrt{\left(\frac{RSS_1+RSS_2}{n_1+n_2-4}\right)\left(\frac{1}{S_{x_1x_1}}+\frac{1}{S_{x_2x_2}}\right)}$$

Interpreting the the confidence interval

If the C.I. for $\beta_1 - \beta_2$ contains 0, we cannot reject the null that (that the two regression lines are parallel) and thus stay with the parallel lines model. It is now natural to fit the parallel lines model and examine it to see whether a even simpler model can be used.

Parallel Lines

Next, we will consider model 2 which assumes two regression lines with equal slopes but different intercept parameters as illustrated in the plot below.



Within the parallel lines model, at any point x, the two lines take the values

$$\alpha_1 + \beta(x - \bar{x}_{1.})$$
$$\alpha_2 + \beta(x - \bar{x}_{2.})$$

and so the difference between the lines is:

$$\begin{aligned}\alpha_1 + \beta(x - \bar{x}_{1.}) - (\alpha_2 + \beta(x - \bar{x}_{2.})) &= \alpha_1 + \beta x - \beta\bar{x}_{1.} - \alpha_2 - \beta x + \beta\bar{x}_{2.} \\ &= \alpha_1 - \alpha_2 + \beta\bar{x}_{2.} - \beta\bar{x}_{1.} \\ &= \alpha_1 - \alpha_2 + \beta(\bar{x}_{2.} - \bar{x}_{1.})\end{aligned}$$

This is simply the distance between the two regression lines (for example the distance between the blue and black lines in the figure above).

We can assess whether a single straight line, with no differences between the groups, model 3, is a suitable model for the data by constructing a C.I. for $\alpha_1 - \alpha_2 + \beta(\bar{x}_{2.} - \bar{x}_{1.})$ and examining whether this interval contains 0.

95% Confidence interval for parallel lines model

Data: (y_{ij}, x_{ij}); $\quad i = 1, 2; \quad j = 1, \ldots, n_i$.

Model : $E(Y_{ij}) = \alpha_i + \beta(x_{ij} - \bar{x}_{i.})$

Calculate the 95% confidence interval (CI) for $\alpha_1 - \alpha_2 + \beta(\bar{x}_{2.} - \bar{x}_{1.})$. The C.I. has the form

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t(n_1+n_2-3;0.975)\sqrt{\frac{RSS}{n_1+n_2-3}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

where $b = \begin{pmatrix} 1 \\ -1 \\ \bar{x}_{2.} - \bar{x}_{1.} \end{pmatrix}$

For this model,

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & (x_{11} - \bar{x}_{1.}) \\ . & . & . \\ . & . & . \\ 1 & 0 & (x_{1n_1} - \bar{x}_{1.}) \\ 0 & 1 & (x_{21} - \bar{x}_{2.}) \\ . & . & . \\ . & . & . \\ 0 & 1 & (x_{2n_2} - \bar{x}_{2.}) \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix}, \qquad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 & 0 \\ 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & \frac{1}{S_{x_1x_1}+S_{x_2x_2}} \end{pmatrix}.$$

For our linear combination, $\mathbf{b}^T\boldsymbol{\beta}$ of interest, we have to spot that

$$\mathbf{b} = \begin{pmatrix} 1 \\ -1 \\ (\bar{x}_{2.} - \bar{x}_{1.}) \end{pmatrix}$$

$\newline

A 95% C.I. for $\mathbf{b}^T\boldsymbol{\beta}$ is

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t(n-p;0.975)\sqrt{\frac{RSS}{n-p}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

$$\hat{\alpha}_1 - \hat{\alpha}_2 + \hat{\beta}(\bar{x}_{2.} - \bar{x}_{1.}) \pm t(n_1+n_2-3;0.975)\sqrt{\frac{RSS}{n_1+n_2-3}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}},$$

where

$$\begin{aligned}
\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b} &= \begin{pmatrix} 1 & -1 & \{\bar{x}_{2.} - \bar{x}_{1.}\} \end{pmatrix} \begin{pmatrix} \frac{1}{n_1} & 0 & 0 \\ 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & \frac{1}{S_{x_1x_1}+S_{x_2x_2}} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ \{\bar{x}_{2.} - \bar{x}_{1.}\} \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{n_1} & -\frac{1}{n_2} & \frac{\{\bar{x}_{2.}-\bar{x}_{1.}\}}{S_{x_1x_1}+S_{x_2x_2}} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ \{\bar{x}_{2.} - \bar{x}_{1.}\} \end{pmatrix} \\
&= \frac{1}{n_2} + \frac{1}{n_2} + \frac{(\bar{x}_{2.} - \bar{x}_{1.})^2}{S_{x_1x_1} + S_{x_2x_2}}.
\end{aligned}$$

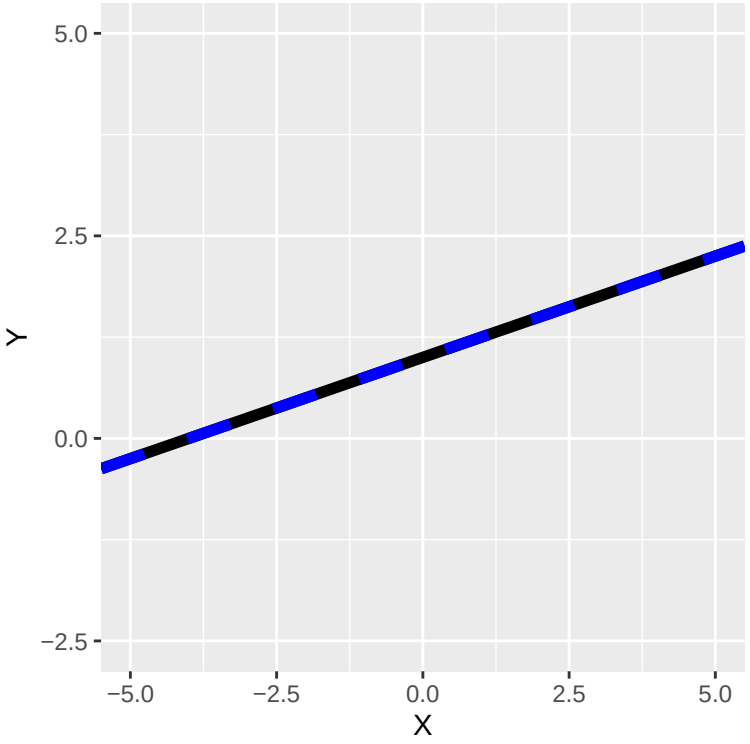
A 95% C.I. for $\mathbf{b}^T\boldsymbol{\beta}$ is therefore

$$\hat{\alpha}_1 - \hat{\alpha}_2 + \hat{\beta}(\bar{x}_{2.} - \bar{x}_{1.}) \pm t(n-p, 0.975)\sqrt{\left(\frac{RSS}{n-p}\right)\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{x}_{2.} - \bar{x}_{1.})^2}{S_{x_1x_1} + S_{x_2x_2}}\right)}$$

$n = n_1 + n_2, p = 3$ and so $n - p = n_1 + n_2 - 3$.

Interperetation of Confidence Interval

If this confidence interval contains 0, we cannot reject the single straight line model, model 3, and thus we stay with the single line model instead of the parallel lines model. A illustrative example is given below. Agian we have two lines in black and blue but we believe these lines to have the same slope and intercept parameters.



ANalysis of COVAriance

The possible models which describe the data are:

$$E(Y_{ij}) = \alpha_i + \beta_i(x_{ij} - \bar{x}_{i.})$$
$$E(Y_{ij}) = \alpha_i + \beta(x_{ij} - \bar{x}_{i.})$$
$$E(Y_{ij}) = \alpha + \beta(x_{ij} - \bar{x}_{..})$$

Some important things to remember are:

1. How to formulate each model in vector-matrix notation as

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

2. How to use the general linear model formulae below to derive the parameter estimates and residual sums of squares as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \quad \text{and} \quad RSS = \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}}$$

3. How to use the general formula for a confidence interval for a linear combination of parameters to compare models.

Additional Reading

Please see

- Section 14.2 in **Linear Models with R**.
- Sections 5.4 in **Regression Analysis By Example**.
- Section 3.3.2 in **An Introduction to Statistical Learning**.