Multiple regression

Bert van der Veen

Department of Mathematical Sciences, NTNU

Introduction

Going to largely omit code in presentation. But see .Rmd files.

Outline Today

- Mutiple linear regression
- Model validation
- Introduction to GLMs

Questions about yesterday?

What if we have >1 explanatory variable?

We often want to look at the impacts of several variables together

- they may all have some effect
- we might be doing an experiment where factors interact
- we might want to model one variable as a polynomial

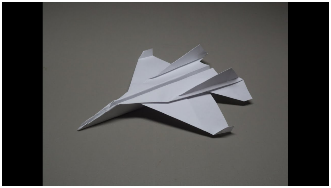
The model

This is our model for simple regression

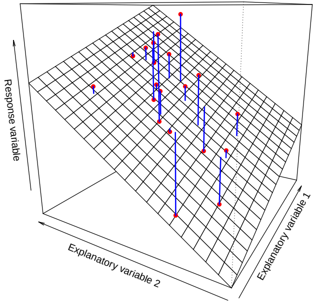
$$y_i = \alpha + \beta x_i + \epsilon_i$$

How can we extend it to more than one variable?

Two explanatory variables

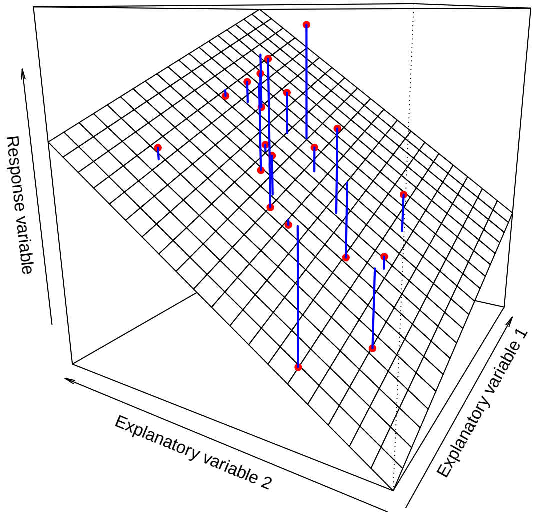


$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$



This is a plane
With more than two covariates it is a **hyperplane**

Plane


Response variable
Explanatory variable 2
Explanatory variable 1

Fitting in R

In R we can just use the same function as we did before.

The only change is in the formula. It was

```
y ~ x
```

now it is

```
y ~ x1 + x2
```

and the same for categorical and continuous covariates.

More than two: general linear regression

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{ik} + \epsilon_i$$

$$y_i = \alpha + \sum_{k=1}^{p} \beta_k x_{ik} + \epsilon_i$$

- we have p covariates, labelled from $k = 1$ to p
- we have p covariate effects
- the k^{th} covariate values for the i^{th} observation is x_{ik}

Design Matrices

We can write this more compactly. First, we turn the intercept into a covariate by filling a column of 1s for every data point. Then we write all of the covariates in a matrix, **X**:

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{pmatrix} 1 & 2.3 & 3.0 \\ 1 & 4.9 & -5.3 \\ 1 & 1.6 & -0.7 \\ \vdots & \vdots & \vdots \\ 1 & 8.4 & 1.2 \end{pmatrix} \end{array}$$

So, the first column is the intercept, and the second and third columns are two covarias.

This is called the *Design Matrix*: it helps to write down the model

Writing the Model

Using matrix algebra, the regression model becomes

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are now all vectors of length n, where there are n data points. $\mathbf{X}$ is am $n \times p$ matrix.

We will not look at the mathematics in any detail: the point here is that the model for the effect of covariates can be written in the design matrix.

The Solution (just so you can see it)

After a bit of matrix algebra, one can find the ML solution:

$$\hat{\mathbf{b}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

where **b** is the MLE for β.

In practice:

- you won't have to calculate this: the computer does it, and
- the computer actually doesn't use this

Writing the Model: continuous covariates

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

is

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 2.3 & 3.0 \\ 1 & 4.9 & -5.3 \\ 1 & 1.6 & -0.7 \\ \vdots & \vdots & \vdots \\ 1 & 8.4 & 1.2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

α is the intercept, β_1 is the slope parameter for x_1, and so on.

Categorical variables

Categorical variables need to be turned into something numerical.

$$\mathbf{x} = \begin{pmatrix} \textbf{Species} \\ \text{Orchid} \\ \text{Orchid} \\ \text{Dandelion} \\ \vdots \\ \text{Daisy} \end{pmatrix} \Rightarrow \mathbf{X} = \begin{pmatrix} \textbf{Orchid} & \textbf{Dandelion} & \textbf{Daisy} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix}$$

But do we need each column?

Contrasts

There are many ways to construct a design matrix for categorical variables.

`constrasts` and `constr.treatment`

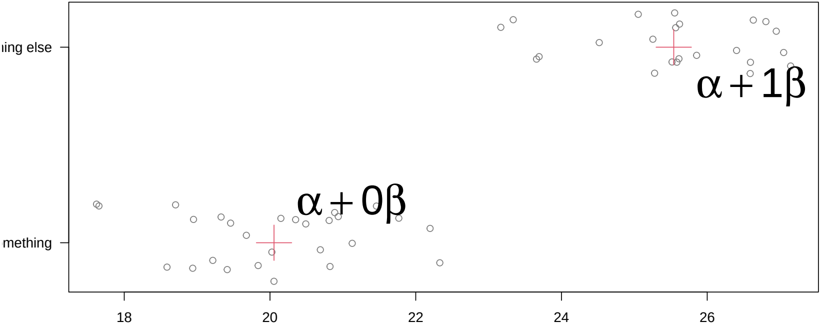
- Treatment contrast are default in `R` (“dummy”)
- Sum-to-zero
- Polynomial
- Difference
- Etc.

Writing the Model: categorical (ANOVA)

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Here, α is the intercept for the first category, β_1 the difference of the first and second category, β_2 the difference between the first and third categories.

Examples of linear models: categorical x_i (from yesterday)



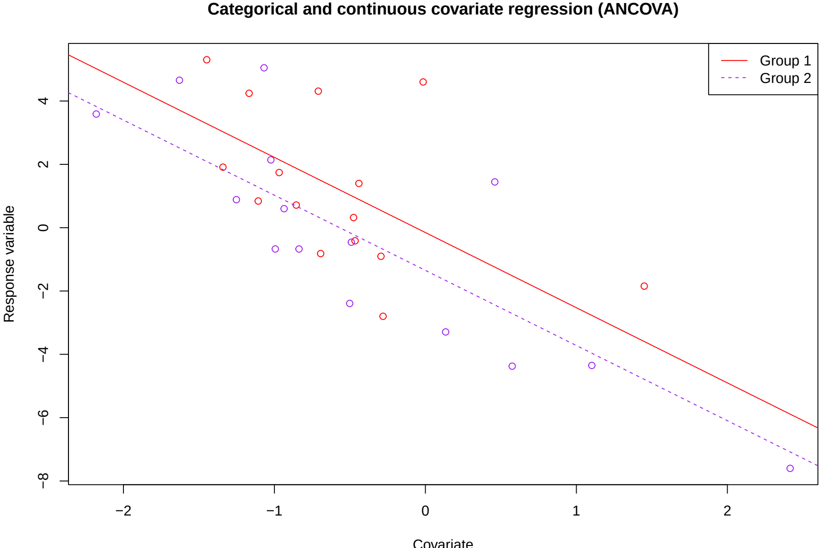
- α is the mean of the first group
- β is the deviation from the mean of the first group

Writing the Model: continuous and categorical

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 3.0 \\ 1 & 1 & -5.3 \\ 1 & 1 & -0.7 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1.2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Here, α is the intercept for the first category at $x_3 = 0$, β_1 is the difference for the second category at $x_3 = 0$, and β_2 is the slope parameter for two regression lines.

Writing the Model: continuous and categorical



Interactions

An interaction is when we have the product of two (or more) covariates in the model:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

`y ~ x1+x2+x1:x2` or `y ~ x1*x2`

It means that we expect the effect of two covariates to jointly impact y_i
It does **not** mean we model how x_1 affect x_2 or vice versa!

Interactions: continuous-continuous

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 2.3 & 3.0 & 2.3 * 3.0 \\ 1 & 4.9 & -5.3 & 4.9 * -5.3 \\ 1 & 1.6 & -0.7 & 1.6 * -0.7 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 8.4 & 1.2 & 8.4 * 1.2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

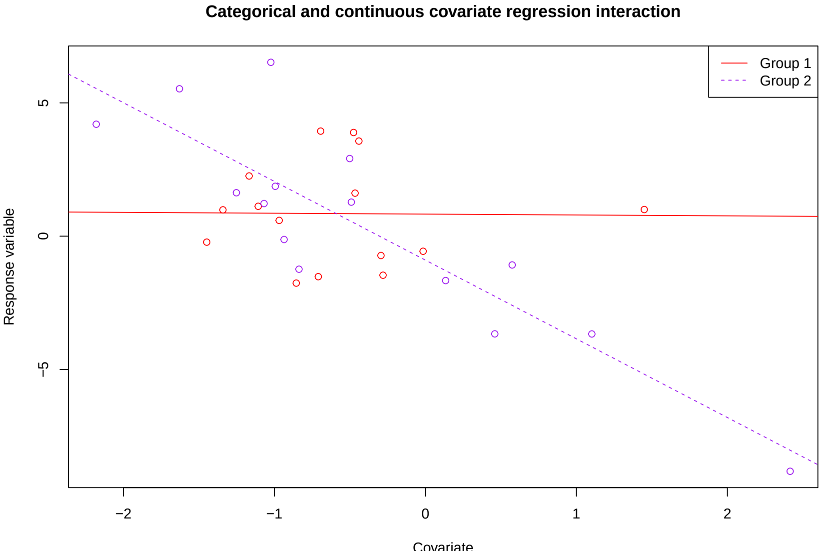
β_1 is the slope for x_2, β_2 is the slope of x_3, β_3 is their joint parameter. It represents how the effect of x_1 or x_2 changes with the other covariate. E.g., water and fertilizer on plant growth.

Interactions: categorical-continuous

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 3.0 & 0 \\ 1 & 1 & -5.3 & -5.3 \\ 1 & 1 & -0.7 & -0.7 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1.2 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

A separate regression line for each category. Here, α and β_2 are the slope and intercept for the regression line of the first category. $\alpha + \beta_1$ is the intercept and $\beta_3 + \beta_4$ is the slope of the regression line for the second category.

Interactions: categorical-continuous

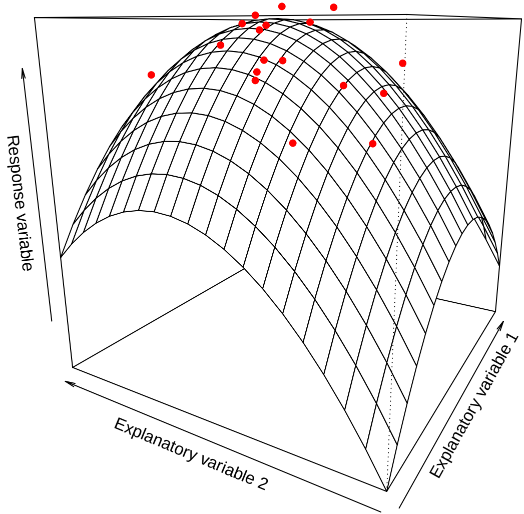


Other functions of explanatory variables

As long as the model is linear in the parameters, we can also have functions:

- Quadratic: $y_i = x_i\beta + x_i^2\beta_2$
- Centering: $y_i = (x_i - \bar{\mathbf{x}})\beta$
- Exponential: $y_i = \exp(x_i)\beta$
  - or logarithmic: $y_i = \log(x_i)\beta$

Surface: quadratic effects



Wiggly things

$$\mathbf{Y} = s(\mathbf{X}) + \epsilon$$

See GAM workshop by Physalia

Finding a "good" model

We do not usually explicitly specify regressions in terms of their imposed hypersurface.

More on how to find a model that fits the data well tomorrow.

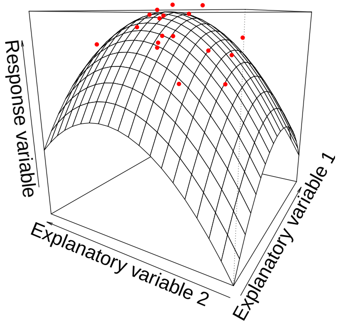
The predict function

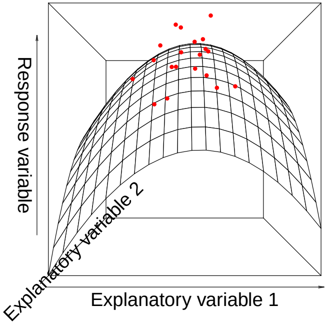
In R we can calulate $\hat{y}_i = \hat{\alpha} + x_i \hat{\beta}_1$ with the `predict` function:

```
predict(model, newdata = newX)
```

Here, `newX` are the values of the covariate that we want to calculate $\hat{y}_i$ for. For the observed values we leave it empty.

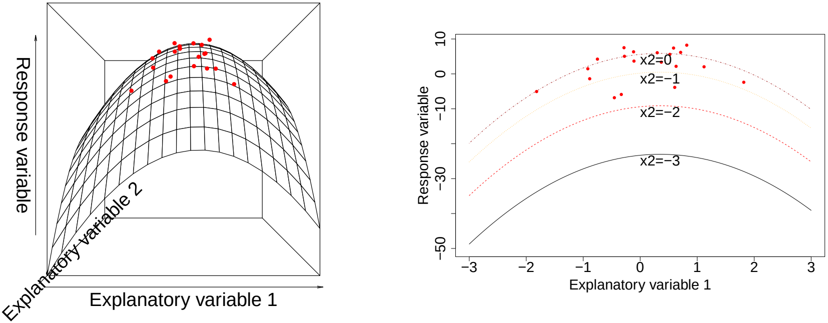
Visualizing a multiple regression





- We want to look at the regression in 2D anyway
- So we need to choose what point to do that from

Visualizing a multiple regression



- We want to look at the regression in 2D anyway
- So we need to choose what point to do that from

Summary

- Multiple regression and the design matrix
- Fortunately we have the `lm()` function in R!

Example code for practical

```
dataset <- read.csv("some_place_on_my_computer/awesomedata.cs
lm(y ~ x1+x2, data = dataset)
```

Questions


IF YOU HAVE
A QUESTION... ASK
YOU WON'T BE
MAULED, PROMISE
imgflip.com