MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: . . . with Data Gathering, 9 hp

Lecture 3b, spring 2023

Multiple linear regression - Categorical x-variables

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

27/3-23

Categorical variables (factors)

- Categorical variables (factors) take a fixed number of non-numerical values, e.g. Male/Female or Red/Blue/Green. There isn't necessarily any logical order between the categories or any obvious translation to numerical values, e.g., "Red $=$ 1, Blue $=$ 2, Green $=$ 3" makes as much sense ($=$ no sense) as "Red $= -14$, Blue $=$ 2.54, Green $=$ 52.4".
- Other times there is some ordering (weight$=\{$underweight, normal, overweight$\}$), however attaching numerical "labels" does not imply admissible mathematical operations. If underweight $=$ 1, normal $=$ 2, overweight $=$ 3 then underweight $+$ normal $=$ 1 $+$ 2 $=$ 3 makes no sense.

Thus they cannot be used as x-variables without some care.

Dummy variables

Create as many new variables as there are categories, e.g., x_{weight} is replaced by the three variables

$$x_{\text{normal}} = \begin{cases} 1 & \text{if } x_{\text{weight}} = \text{normal}, \\ 0 & \text{otherwise} \end{cases}$$

$$x_{\text{under}} = \begin{cases} 1 & \text{if } x_{\text{weight}} = \text{underweight}, \\ 0 & \text{otherwise} \end{cases}$$

$$x_{\text{over}} = \begin{cases} 1 & \text{if } x_{\text{weight}} = \text{overweight}, \\ 0 & \text{otherwise} \end{cases}$$

x_{weight}	x_{normal}	x_{under}	x_{over}
normal	1	0	0
underweight	0	1	0
overweight	0	0	1

The model $Y_i = \beta_0 + \beta_{\text{weight}} x_{i,\text{weight}} + \epsilon_i$ using dummy-variables would then be expressed as:
$Y_i = \beta_0 + \beta_{\text{normal}} x_{\text{i,normal}} + \beta_{\text{under}} x_{\text{i,under}} + \beta_{\text{over}} x_{\text{i,over}} + \epsilon_i$.

Problem!

The matrix $\mathbf{X}'\mathbf{X}$ with

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,\text{normal}} & x_{1,\text{under}} & x_{1,\text{over}} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,\text{normal}} & x_{n,\text{under}} & x_{n,\text{over}} \end{pmatrix} = [\text{e.g.}] = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

is singular and cannot be inverted! This is because, for any row in $\mathbf{X}$, we have $x_{i,\text{normal}} + x_{i,\text{under}} + x_{i,\text{over}} \equiv 1 =$ the value in the first column of $\mathbf{X}$.

Hence $\mathbf{X}$ does not have *full rank*, as one of the columns can be determined from the others (→ redundant information).

Solution

Remove one of the columns (this does not imply any loss of information)

(a) Delete the intercept:
$Y_i = \beta'_{\text{normal}} x_{i,\text{normal}} + \beta'_{\text{under}} x_{i,\text{under}} + \beta'_{\text{over}} x_{i,\text{over}} + \epsilon_i$.

(b) Delete one of the categories (e.g. delete "normal"):
$Y_i = \beta_0 + \beta_{\text{under}} x_{i,\text{under}} + \beta_{\text{over}} x_{i,\text{over}} + \epsilon_i$.

$$\begin{aligned}\beta'_{\text{normal}} &= \beta_0 \\ \beta'_{\text{under}} &= \beta_0 + \beta_{\text{under}} \\ \beta'_{\text{over}} &= \beta_0 + \beta_{\text{over}}\end{aligned}$$

Solution (b) implies that:

- "normal" is the reference category or baseline
- the intercept is the expected response for the reference category
- parameters for the other categories give the category systematic effect in relation to the reference category.

Example

For an underweight subject

$$E(Y \mid \text{weight} = \text{under}) = \beta_0 + \beta_{\text{under}} = \beta_{\text{normal}} + \beta_{\text{under}}$$

- The expected outcome for an underweight subject is $\beta_0 \equiv \beta_{\text{normal}}$ plus an increment (positive or negative) β_{under}.
- That's why "normal" is said to be a reference category and β_{under} a differential effect.

Warning

When (b) is used (creation of reference category) do not interpret parameters as when you have continuous covariates! Parameters for categories have to be interpreted in relation to the reference category.

Which category to choose as reference?

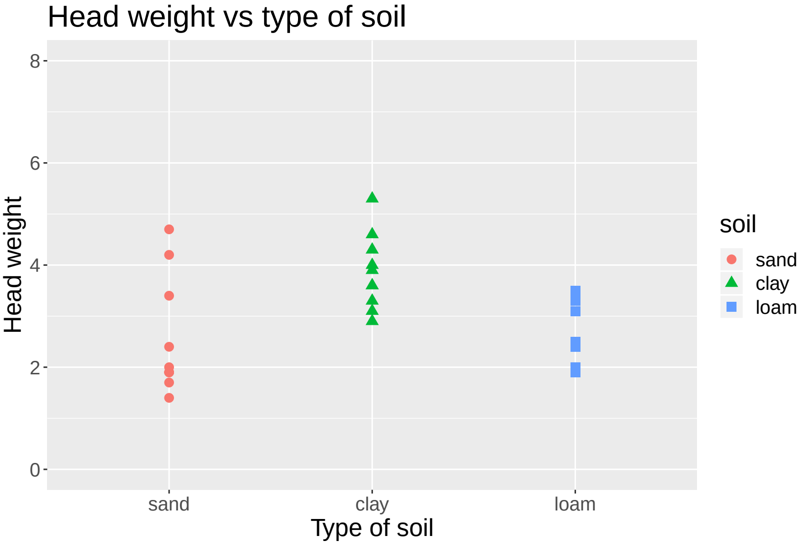
This is problem specific: say the one which makes more sense to be used as a term of comparison. In some contexts it is natural/obvious which one to consider as a "normal" or "default" level.
However, if the number of observations in the reference category is small, all β-estimates will be uncertain! The reference category should always be large.

R calls categorical variables `factors`.
Categories for a categorical variable are called `levels`.

Example: Cabbage

In an agricultural experiment we have grown cabbages in three different types of soil: sand, clay and loam. We have also used different ammounts of fertilizer. We want to model their effect on the weight of the cabbage heads.

soil	fert.	headwt	soil	fert.	headwt	soil	fert.	headwt
sand	10	1.4	clay	10	3.1	loam	10	1.9
sand	15	1.9	clay	15	2.9	loam	15	2.0
sand	20	3.4	clay	20	3.6	loam	20	3.1
sand	25	2.4	clay	25	3.9	loam	25	3.5
sand	30	4.2	clay	30	3.6	loam	30	3.3
sand	35	1.9	clay	35	4.0	loam	35	2.4
sand	40	1.7	clay	40	3.3	loam	40	2.5
sand	45	4.7	clay	45	4.3	loam	45	3.5
sand	50	1.9	clay	50	5.3	loam	50	1.9
sand	55	2.0	clay	55	4.6	loam	55	3.3



observed data

Cabbage: soil model and estimates

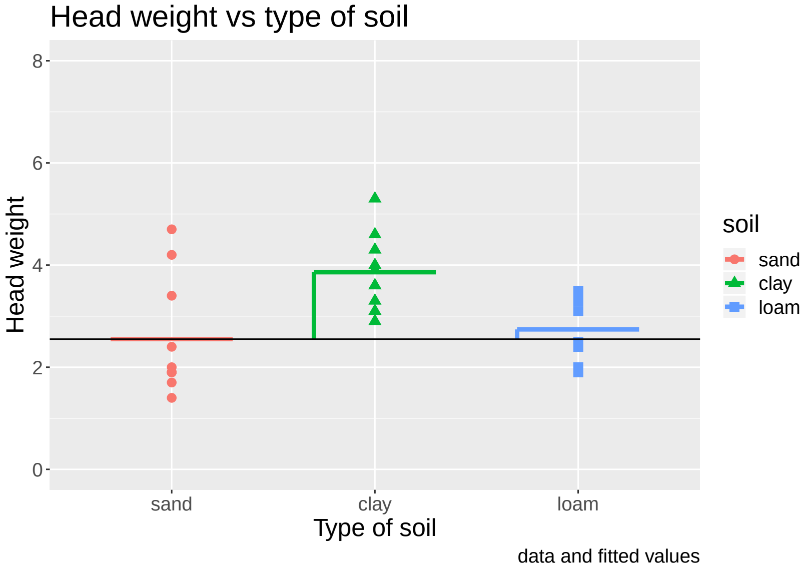
$Y =$ head weight, $x_1 = \begin{cases} 1 & \text{clay} \\ 0 & \text{not clay} \end{cases}$, $x_2 = \begin{cases} 1 & \text{loam} \\ 0 & \text{not loam} \end{cases}$

Model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \epsilon_i & \text{sand} \\ \beta_0 + \beta_1 + \epsilon_i & \text{clay} \\ \beta_0 + \beta_2 + \epsilon_i & \text{loam} \end{cases}$

Variable	parameter	estimate	s.e.	95 % C.I.
intercept (sand)	β_0	2.55	0.28	$(1.98, 3.12)$
clay (vs sand)	β_1	1.31	0.39	$(0.51, 2.11)$
loam (vs sand)	β_2	0.19	0.39	$(-0.61, 0.99)$
resid.std.dev	σ	0.87	df $= 27$	

Fitted “line”:

$$\hat{Y} = 2.55 + 1.31x_1 + 0.19x_2 = \begin{cases} 2.55 & \text{sand} \\ 2.55 + 1.31 & \text{clay} \\ 2.55 + 0.19 & \text{loam} \end{cases}$$


Head weight vs type of soil
Head weight
8
6
4
2
0
sand
clay
loam
Type of soil
soil
sand
clay
loam
data and fitted values

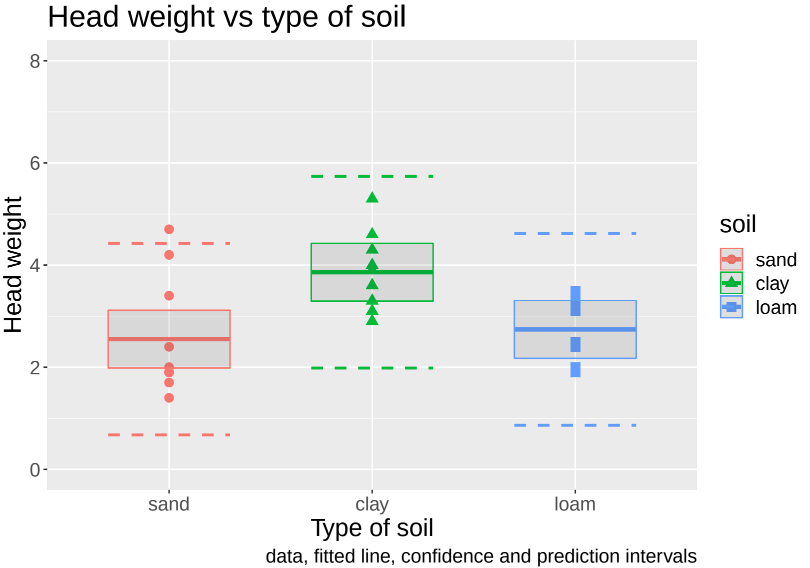
Predictions

What is the average head weight for the different soil types?
What head weights might we observe?

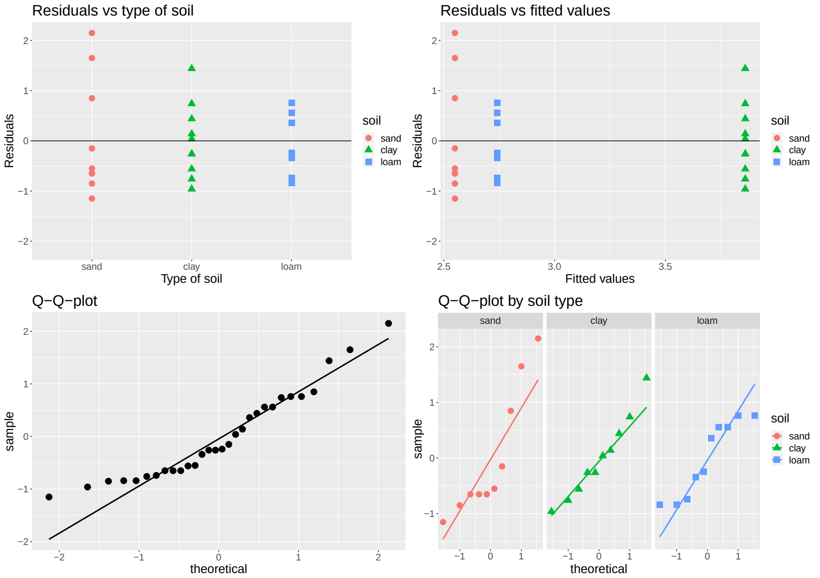
on average	estimate	s.e.	95 % C.I.
soil	$\hat{Y}_{\text{soil}} = \hat{\beta}_0 = 2.55$	0.28	$(1.98, 3.12)$
clay	$\hat{Y}_{\text{clay}} = \hat{\beta}_0 + \hat{\beta}_1 = 3.86$	0.28	$(3.29, 4.43)$
loam	$\hat{Y}_{\text{loam}} = \hat{\beta}_0 + \hat{\beta}_2 = 2.74$	0.28	$(2.17, 3.31)$

single head	estimate	s.e.	95 % P.I.
soil	$\hat{Y}_{\text{pred}_{\text{soil}}} = 2.55 + \epsilon_0$	0.91	$(0.67, 4.43)$
clay	$\hat{Y}_{\text{pred}_{\text{clay}}} = 3.86 + \epsilon_0$	0.91	$(1.98, 5.74)$
loam	$\hat{Y}_{\text{pred}_{\text{loam}}} = 2.74 + \epsilon_0$	0.91	$(0.86, 4.62)$

Note: the different soil types have the same standard errors here because they have the same number of observations.


Head weight vs type of soil
Head weight
8
6
4
2
0
sand
clay
loam
Type of soil
soil
sand
clay
loam
data, fitted line, confidence and prediction intervals

Basic residual analysis



Residuals vs type of soil

Residuals vs fitted values

Q–Q–plot

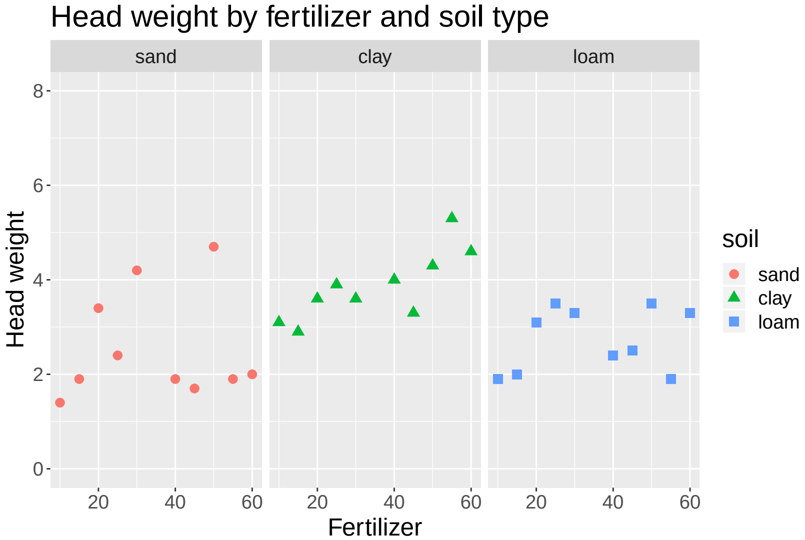
Q–Q–plot by soil type

Conclusions

- The difference in average head weight between sand and loan is not necessarily different from 0 (the confidence interval $I_{\beta_2} = (-0.61, 0.99)$ contains 0).
- The residual variability in loam seems smaller than in the other two soil types.
- The residuals are not very un-normal.

Questions

- Would adding the amount of fertilizer to the model improve the fit?
- Is the effect of the fertilizer the same for all soil types?


Head weight by fertilizer and soil type
sand
clay
loam
Head weight
0
2
4
6
8
20
40
60
Fertilizer
soil
sand
clay
loam

data

Cabbage: soil and fertilizer model

$Y =$ head weight, $x_1 = \begin{cases} 1 & \text{clay} \\ 0 & \text{not clay} \end{cases}$, $x_2 = \begin{cases} 1 & \text{loam} \\ 0 & \text{not loam} \end{cases}$, $x_3 =$ fertilizer. Model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_3 + \epsilon_i \\ &= \begin{cases} \beta_0 + \beta_3 x_3 + \epsilon_i & \text{sand} \\ \beta_0 + \beta_1 + \beta_3 x_3 + \epsilon_i & \text{clay} \\ \beta_0 + \beta_2 + \beta_3 x_3 + \epsilon_i & \text{loam} \end{cases} \qquad \epsilon_i \sim N(0, \sigma^2) \end{aligned}$$

Note: This is three parallel lines with the same fertilizer-slope but different intercepts for the different soil types.

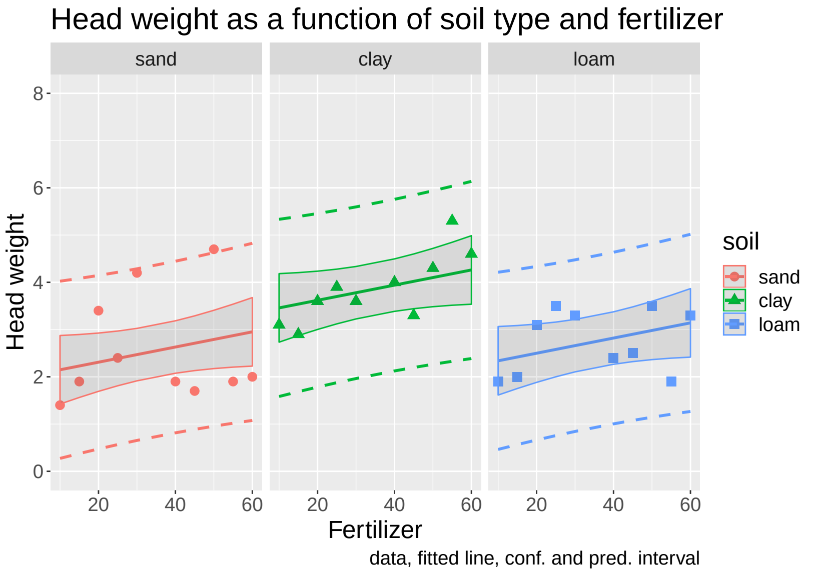
Cabbage: soil and fertilizer estimates

Variable	parameter	estimate	s.e.	95 % C.I.
intercept (sand)	β_0	1.99	0.42	$(1.13, 2.85)$
clay (vs sand)	β_1	1.31	0.38	$(0.54, 2.08)$
loam (vs sand)	β_2	0.19	0.38	$(-0.58, 0.96)$
fertilize	β_3	0.016	0.009	$(-0.003, 0.035)$
resid.std.dev	σ	0.84	df $= 26$	

Note: The confidence interval for β_3 covers 0. It is possible that fertilizer has no effect. Also, loam might not be different from sand.

Fitted lines:

$$\hat{Y} = 1.99 + 1.31x_1 + 0.19x_2 + 0.0016x_3$$
$$= \begin{cases} 1.99 + 0.0016x_3 & \text{sand} \\ 1.99 + 1.31 + 0.0016x_3 & \text{clay} \\ 1.99 + 0.19 + 0.0016x_3 & \text{loam} \end{cases}$$


Head weight as a function of soil type and fertilizer
sand
clay
loam
Head weight
0
2
4
6
8
20
40
60
Fertilizer
soil
sand
clay
loam
data, fitted line, conf. and pred. interval

Different effect of fertilizer on different soil types: interaction

$Y =$ head weight, $x_1 = \begin{cases} 1 & \text{clay} \\ 0 & \text{not clay} \end{cases}$, $x_2 = \begin{cases} 1 & \text{loam} \\ 0 & \text{not loam} \end{cases}$,
$x_3 =$ fertilizer.
Model (note that we get two interaction terms):

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon_i \\ &= \begin{cases} \beta_0 + \beta_3 x_3 + \epsilon_i & \text{sand} \\ \beta_0 + \beta_1 + (\beta_3 + \beta_4) x_3 + \epsilon_i & \text{clay} \\ \beta_0 + \beta_2 + (\beta_3 + \beta_5) x_3 + \epsilon_i & \text{loam} \end{cases} \qquad \epsilon_i \sim N(0, \sigma^2) \end{aligned}$$

Note: The extra parameters β_4 and β_5 signify how the fertilizer slope for sand should be adjusted to become the fertilizer slope for clay and loam, respectively. If $\beta_4 = 0$ then clay has the same slope as sand. If $\beta_5 = 0$ then loam has the same slope as sand.

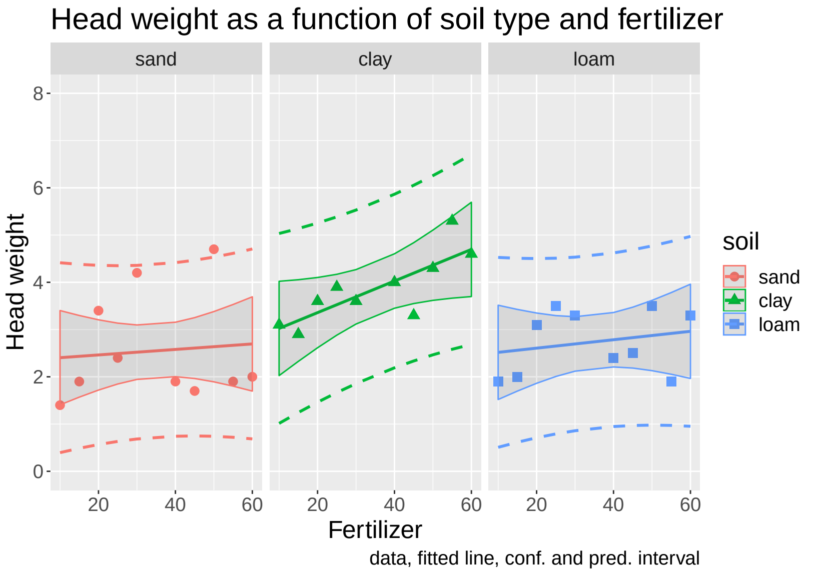
Cabbage: soil and fertilizer interaction: estimates

Variable	parameter	estimate	s.e.	95 % C.I.
intercept (sand)	β_0	2.35	0.62	$(1.06, 3.63)$
clay	β_1	0.34	0.88	$(-1.48, 2.16)$
loam	β_2	0.08	0.88	$(-1.74, 1.90)$
fertilize (sand)	β_3	0.0058	0.0016	$(-0.027, 0.039)$
fert:clay	β_4	0.028	0.023	$(-0.019, 0.075)$
fert:loam	β_5	0.003	0.023	$(-0.044, 0.075)$
resid.std.dev	σ	0.85	df $= 24$	

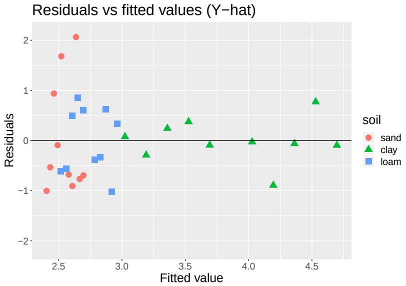
Fitted lines:

$$\hat{Y} = 2.35 + 0.34x_1 + 0.08x_2 + 0.0058x_3 + 0.028x_1x_3 + 0.003x_2x_3$$

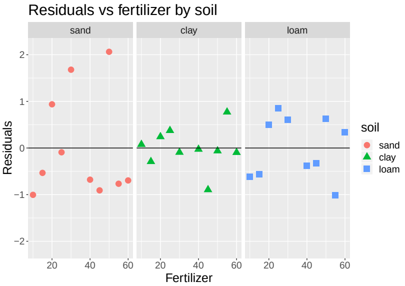
$$= \begin{cases} 2.35 + 0.0058x_3 & \text{sand} \\ 2.35 + 0.34 + (0.0058 + 0.028)x_3 & \text{clay} \\ 2.35 + 0.08 + (0.0058 + 0.003)x_3 & \text{loam} \end{cases}$$


Head weight as a function of soil type and fertilizer
sand
clay
loam
Head weight
0
2
4
6
8
20
40
60
Fertilizer
soil
sand
clay
loam
data, fitted line, conf. and pred. interval

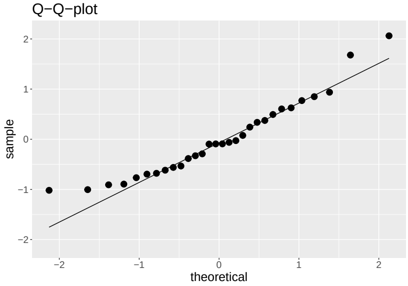
Residuals vs fitted values (Y−hat)



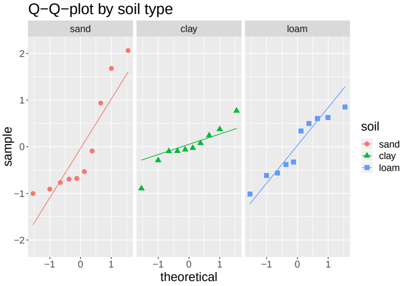
Residuals vs fertilizer by soil



Q−Q−plot



Q−Q−plot by soil type



Some conclusions from the residual analysis

- The residual variability in clay is now as small as that for loam! The fertilizer explained most of it.
- The residual variability in sand is still large and un-explained.
- The residuals are slightly closer to a normal distribution.

Questions for the future (next lecture)

- We have only 30 observations to estimate the six β-parameters. This gives uncertain estimates...
- ...and so, almost all of the β-parameters have confidence intervals that cover zero. Are there some variables that could/should be removed from the model?
- How can we test whether we need, e.g., the interaction terms?