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

Lecture 3, spring 2019

Introduction to multiple linear regression; Categorical variables

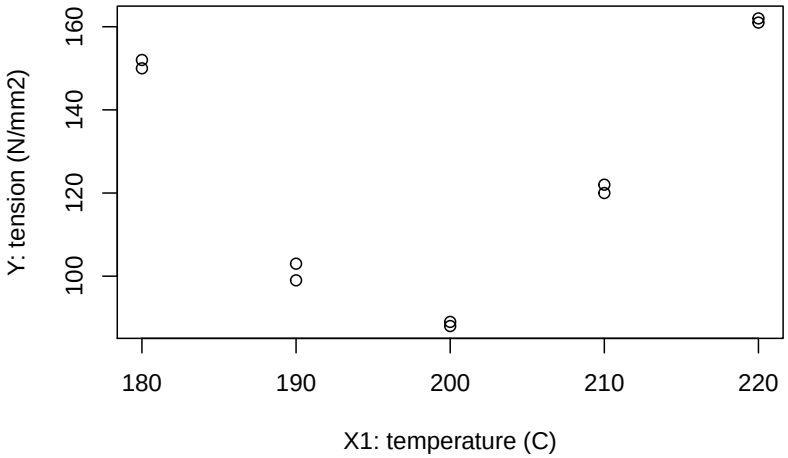
Mathematical Statistics / Centre for Mathematical Sciences
Lund University

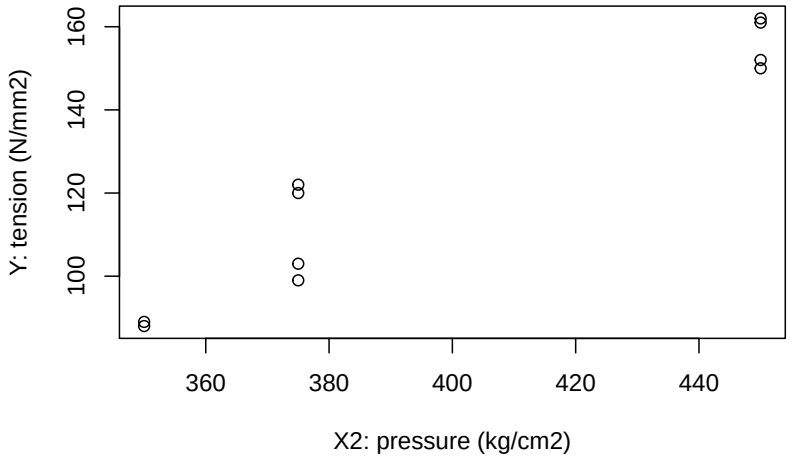
1/4-19

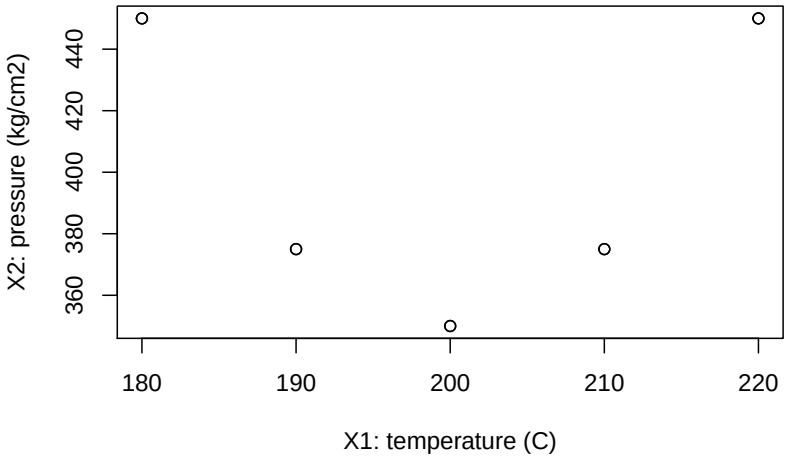
Multiple regression: Example

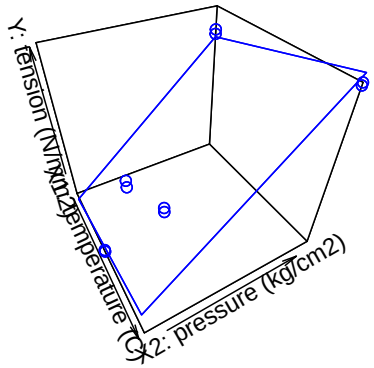
Module of elasticity as a function of pressure and temperature:
Temperature and pressure and resulting tension in 10 plastic parts;

Tension (Y) (N/mm^2)	Temperature (X_1) ($^\circ C$)	Pressure (X_2) (kg/cm^2)
152	180	450
150	180	450
103	190	375
99	190	375
88	200	350
89	200	350
122	210	375
120	210	375
162	220	450
161	220	450


Y: tension (N/mm2)
100
120
140
160
180
190
200
210
220
X1: temperature (C)


Y: tension (N/mm2)
100
120
140
160
360
380
400
420
440
X2: pressure (kg/cm2)


X2: pressure (kg/cm2)
360
380
400
420
440
180
190
200
210
220
X1: temperature (C)


Y: tension (N/mm2)
X1: temperature (C)
X2: pressure (kg/cm2)

From the previous plots what we should deduce is:

- plots of Y vs individual covariates only unveil *partial relationships*. We do not know what happens when other covariates vary together.
- we can discover pairwise relationships between covariates by plotting X_1 vs X_2
- plotting X_1 vs X_2 does not say anything about the 3D joint relationship of (X_1, X_2, Y)
- if the plot (X_1, Y) is nonlinear, you can perhaps transform X_1 and/or Y but again, this is only going to linearize a partial relationship...
- our model (next slide) is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \qquad (*)$$

and in this case even if some **partial** relationships (X_j, Y) are nonlinear, the **entire** surface (*) above is perfectly suitable for the joint relationship.

Multiple linear regression model

$$Y_i = \underbrace{\beta_0}_{\beta_0 \cdot 1} + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i \text{ for } i = 1, \ldots, n$$

where $\epsilon_i \sim N(0, \sigma^2)$ are independent.

Matrix formulation

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1p} \\ 1 & X_{21} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{pmatrix} 1 \cdot \beta_0 + X_{11} \cdot \beta_1 + \cdots + X_{1p} \cdot \beta_p \\ 1 \cdot \beta_0 + X_{21} \cdot \beta_1 + \cdots + X_{2p} \cdot \beta_p \\ \vdots \\ 1 \cdot \beta_0 + X_{n1} \cdot \beta_1 + \cdots + X_{np} \cdot \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}) \quad (n\text{-dimensional multivariate normal distribution})$$

We have $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ where $E(\boldsymbol{\epsilon}) = \mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ and

$$Var(\boldsymbol{\epsilon}) = \begin{pmatrix} V(\epsilon_1) & Cov(\epsilon_1, \epsilon_2) & \dots & Cov(\epsilon_1, \epsilon_n) \\ Cov(\epsilon_2, \epsilon_1) & V(\epsilon_2) & \dots & Cov(\epsilon_2, \epsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\epsilon_n, \epsilon_1) & Cov(\epsilon_n, \epsilon_2) & \dots & V(\epsilon_n) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 \cdot \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \sigma^2\mathbf{I}$$

- A given β_j expresses the *effect* of a change in covariate X_j on the expected value of Y, *given all other covariates in the model*;
- that is, β_j gives the change in $E(Y)$ when X_j increases by 1 units, when all other covariates are kept fixed.
- in other words, β_j can only represent the partial (marginal) effect of X_j on Y; the effect is *conditional* on what other variables we have in the model.
- The relevance of X_j (hence the relevance of β_j) can be different if we introduce other covariates in the model.

The latter two concepts will be emphasized when we talk about hypothesis tests.

We want to find the vector $\hat{\boldsymbol{\beta}}$ that minimizes

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_p X_{ip})^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{e}'\boldsymbol{e}$$

Normal equations

The solution satisfies: $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$

$$\begin{pmatrix} n & \sum_{i=1}^n X_{i1} & \cdots & \sum_{i=1}^n X_{ip} \\ \sum_{i=1}^n X_{i1} & \sum_{i=1}^n X_{i1}^2 & \cdots & \sum_{i=1}^n X_{i1}X_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n X_{ip} & \sum_{i=1}^n X_{i1}X_{ip} & \cdots & \sum_{i=1}^n X_{ip}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_{i1}Y_i \\ \vdots \\ \sum_{i=1}^n X_{ip}Y_i \end{pmatrix}$$

Estimated parameters: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
Uniqueness of $\hat{\boldsymbol{\beta}}$: exist only if $(\mathbf{X}'\mathbf{X})^{-1}$ exists!
Estimated plane: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$
Residuals: $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ = observed $-$ predicted
Estimated variance: $\hat{\sigma}^2 = s^2 = \dfrac{\sum_{i=1}^n e_i^2}{n-(p+1)} = \dfrac{\mathbf{e}'\mathbf{e}}{n-(p+1)}$

Properties of parameter estimates

$$E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}_{\mathbf{I}}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var(\mathbf{Y})\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \sigma^2\mathbf{I} \cdot \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

(Using $Var(\mathrm{AW}) = \mathrm{A}\,Var(\mathrm{W})\mathrm{A}'$ for a constant matrix A)

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) \quad \text{(multivariate normal)}$$

$$\hat{Y}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}} \sim N(\mathbf{x}_0\boldsymbol{\beta}, \sigma^2\mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0') \quad \text{(1D normal)}$$

$$\hat{Y}_{\text{pred}_0} = \mathbf{x}_0\hat{\boldsymbol{\beta}} + \epsilon_0 \sim N(\mathbf{x}_0\boldsymbol{\beta}, \sigma^2(1 + \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0')) \quad \text{(1D normal)}$$

where $\mathbf{x}_0 = \begin{pmatrix} 1 & x_{01} & \dots & x_{0p} \end{pmatrix}$.

Confidence intervals

A $(1-\alpha)\%$ confidence interval for β_j:

$$I_{\beta_j} = \left(\hat{\beta}_j \pm t_{\alpha/2, n-(p+1)} \cdot s \sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}} \right)$$

where $(\mathbf{X}'\mathbf{X})^{-1}_{jj}$ is the j:th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ for $j = 0, 1, \ldots, p$.

A $(1-\alpha)\%$ confidence interval for the expected value $E(Y_0) = \mathbf{x}_0\boldsymbol{\beta}$:

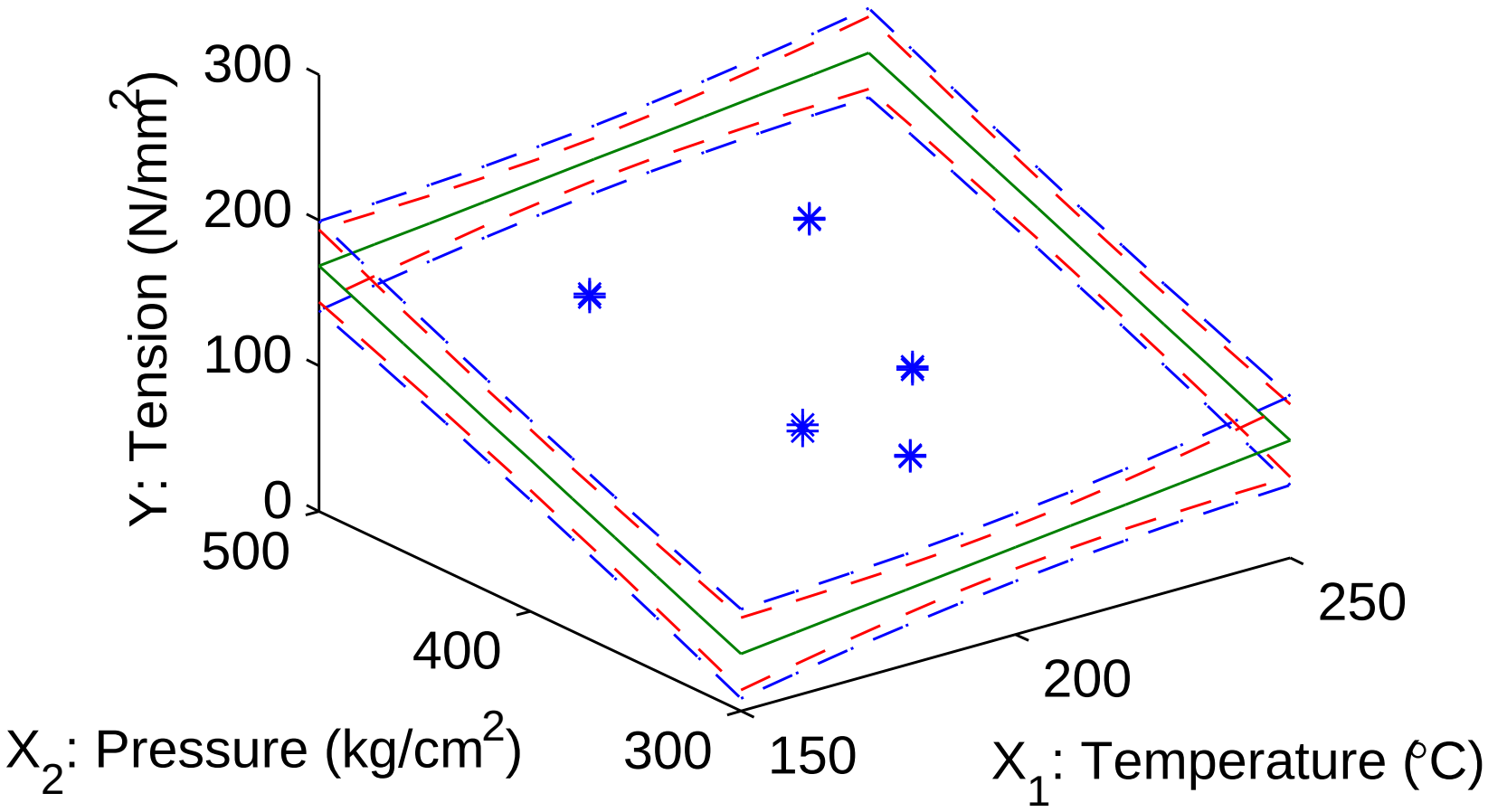
$$I_{E(Y_0)} = \left(\mathbf{x}_0\hat{\boldsymbol{\beta}} \pm t_{\alpha/2, n-(p+1)} \cdot s \sqrt{\mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'} \right)$$

A $(1-\alpha)\%$ prediction interval for a *future response* $Y_{\text{pred}_0} = \mathbf{x}_0\boldsymbol{\beta} + \epsilon_0$:

$$I_{Y_{\text{pred}_0}} = \left(\mathbf{x}_0\hat{\boldsymbol{\beta}} \pm t_{\alpha/2, n-(p+1)} \cdot s \sqrt{1 + \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'} \right)$$

Intervals for the plane

Estimated plane (green); confidence interval for the plane (red) and prediction interval for observations (blue).



t-test in multiple regression

Same formulas as for simple regression: but need some care in the interpretation. Assume we have $p+1$ β-parameters.

$$T = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}}$$

$H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ reject H_0 at significance level α if

- $|T| > t_{\alpha/2,n-(p+1)}$ or $0 \notin I_{\beta_j}$ or P-value $< \alpha$.

In multiple regression it tests the relevance of covariate X_j (in "explaining" $E(Y)$) **given all other covariates in the model**. Individual significance of X_j might change if we add/remove other variables in the model. Significance is relative!

other tests...

Other tests will be introduced next lecture: for example

- can we check the global aptness of the model? We will test simultaneously the **whole** vector $\boldsymbol{\beta}$.
- can we check whether sub-blocks of parameters are relevant? (i.e. does a smaller model provide enough explanation?).

Collinearity (re-discussed later in the course)

- in order to determine $\hat{\beta}$ the matrix $\mathbf{X}'\mathbf{X}$ must be invertible.
- The matrix $\mathbf{X}$ is singular and $(\mathbf{X}'\mathbf{X})^{-1}$ has no unique solution if a linear combination of some of the X-variables equals one of the other X-variables. β cannot be uniquely estimated.
- The matrix $\mathbf{X}$ is nearly singular and $(\mathbf{X}'\mathbf{X})^{-1}$ has an unstable solution if a linear combination of some of the X-variables almost equals one of the other X-variables. The same (almost) information included in several variables.
- β-estimates will have huge variance.
- Correlated X-variables "compete" (one variable might be significant if the other is not in the model, but not if both are in the model, etc.)
- Found by: plotting all X-variables against each other. (R: use `pairs()`)
- Solution: Use only one of the problematic variables

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 (=none) 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

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

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

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

X_{weight}	X_{normal}	X_{under}	X_{over}
normal	1	0	0
under	0	1	0
over	0	0	1

The model $Y_i = \beta_0 + \beta_1 X_{\text{weight,i}} + \epsilon_i$ using dummy-variables would then be expressed as:

$$Y_i = \beta_0 + \beta_{\text{normal}} X_{\text{normal,i}} + \beta_{\text{under}} X_{\text{under,i}} + \beta_{\text{over}} X_{\text{over,i}} + \epsilon_i$$

.

Example:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ \vdots & & & \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Problem!

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

$$\mathbf{X} = \begin{pmatrix} 1 & X_{\text{normal},1} & X_{\text{under},1} & X_{\text{over},1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{\text{normal},n} & X_{\text{under},n} & X_{\text{over},n} \end{pmatrix}$$

is singular and cannot be inverted because for any row in $\mathbf{X}$ $X_{\text{normal}} + X_{\text{under}} + X_{\text{over}} \equiv 1$. The sum returns 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 ($\rightarrow$ redundant information).

Solution

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

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

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

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

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 effect **in relation to the reference category**.

Example

For an underweight subject

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

- Therefore the expected outcome for an under weight 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 "baseline"?
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 relerence group should be a large group.

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

Example

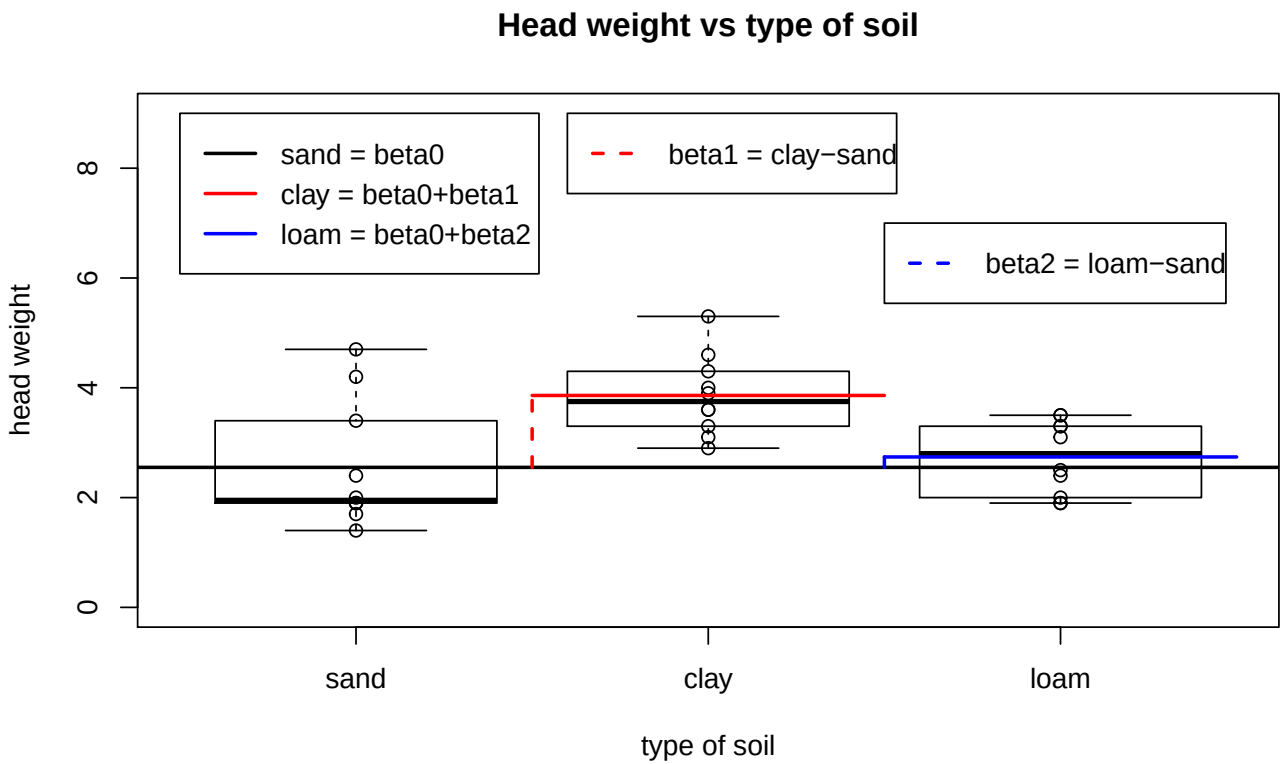
In an agricultural experiment we have grown cabbages in three different types of soil: sand, clay and loam. We have also used different ammounts of fertilizer. We want to model their effect on the weight of the cabbage heads.

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

Model with sand as reference category:

$Y_i = \beta_0 + \beta_1 X_{\text{clay},i} + \beta_2 X_{\text{loam},i} + \epsilon_i$

Estimates: $\hat{\beta}_0 = 2.55$; $\hat{\beta}_1 = 1.31$; $\hat{\beta}_2 = 0.19$.



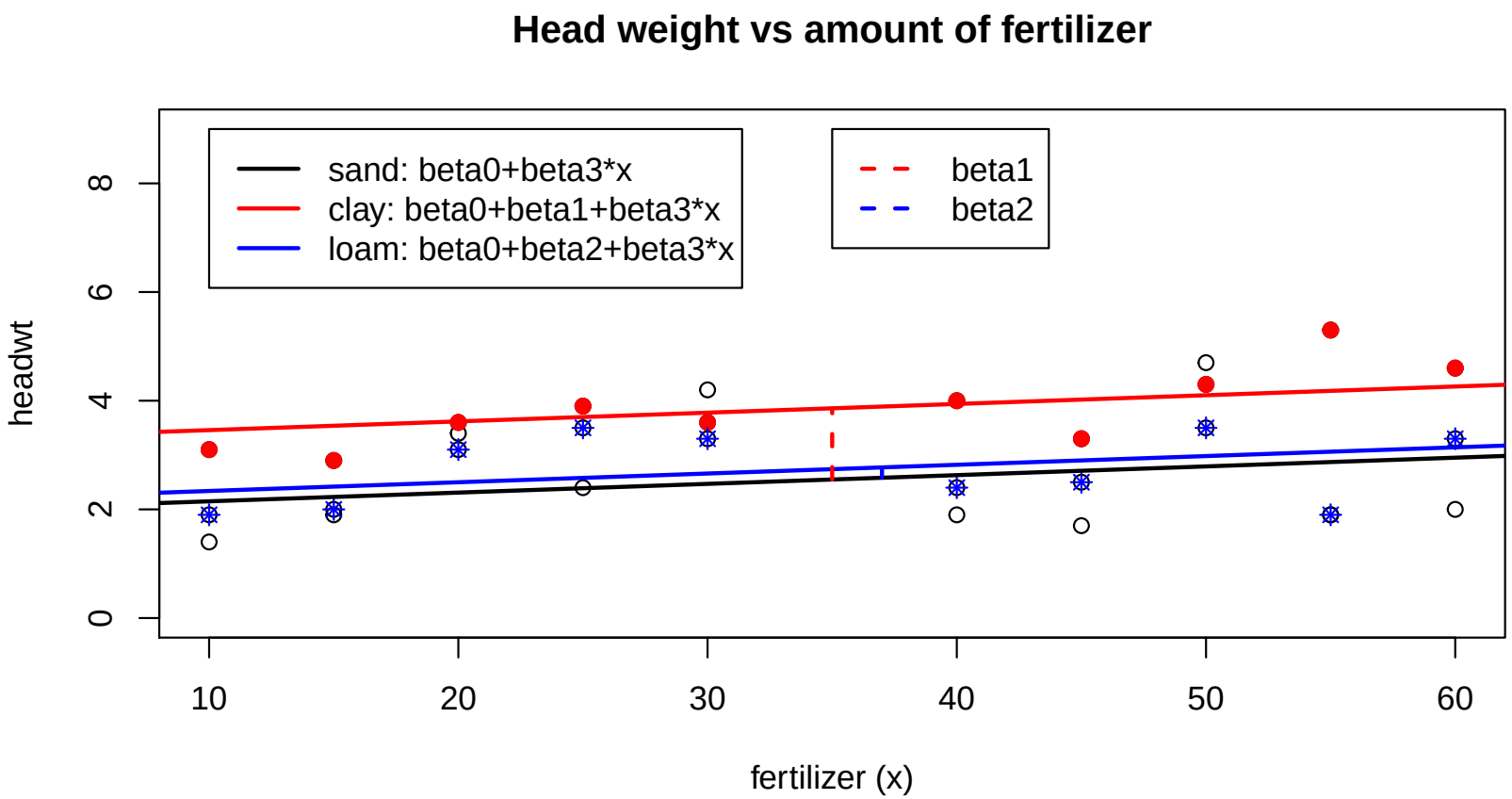
$\hat{\beta}_2$ is not significant. This means that cabbages grown on loam could have the same average head weight as those grown on sand.

One categorical and one continuous X-variable

One slope with different intercepts:

The model $Y_i = \beta_0 + \beta_1 X_{\text{clay},i} + \beta_2 X_{\text{loam},i} + \beta_3 X_3 + \epsilon_i$.

The estimates become $\hat{\beta}_0 = 1.99$; $\hat{\beta}_1 = 1.31$; $\hat{\beta}_2 = 0.19$; $\hat{\beta}_3 = 0.016$.



Interaction

Up to now we have implicitly assumed the *effect* of a given covariate on the response (measured by estimating its β_j) as a constant. Can this effect actually vary according to variation in other covariates?
Example: does the effect of the fertilizer (X_3) vary according to the type of soil (X_{clay}, X_{loam})?

An interaction occurs when the magnitude of the effect of one independent variable X_1 on Y varies as a function of a second independent variable X_2.

Basically when there is a significant interaction the simultaneous influence of two covariates on the response is **not simply additive**.

Interaction

Different intercepts *and* slopes for different types:

$$Y = \beta_0 + \beta_1 X_{\text{clay}} + \beta_2 X_{\text{loam}} + \beta_3 X_3 + \beta_4 X_{\text{clay}} X_3 + \beta_5 X_{\text{loam}} X_3 + \epsilon$$

$$= \begin{cases} \beta_0 + \beta_3 X_3 + \epsilon & \text{if sand,} \\ \beta_0 + \beta_1 + (\beta_3 + \beta_4) X_3 + \epsilon & \text{if clay,} \\ \beta_0 + \beta_2 + (\beta_3 + \beta_5) X_3 + \epsilon & \text{if loam,} \end{cases}$$