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

Lecture 3, spring 2024

Multiple linear regression - collinearity and interaction - categorical x-variables

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

25/3-24

Multiple linear regression model

See Lecture 1+2 for details.

- $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$ for $i = 1, \ldots, n$ where $\epsilon_i \sim N(0, \sigma^2)$ are independent.
- $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

Estimates

- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})$
- Fitted values: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})$
- Residuals: $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} =$ observed $-$ predicted
- Residual variance: $\hat{\sigma}^2 = s^2 = \dfrac{\sum_{i=1}^n e_i^2}{n-(p+1)} = \dfrac{\mathbf{e}^\mathsf{T}\mathbf{e}}{n-(p+1)}$

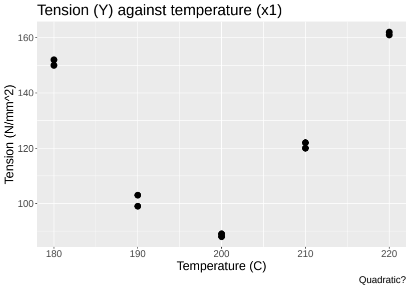
- A given β_j expresses the *effect* of a change in covariate x_j on the expected value of Y, *given all other covariates in the model*;
- that is, β_j gives the change in $E(Y)$ when x_j increases by 1 unit, when all other covariates are kept fixed.
- in other words, β_j can only represent the partial (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.

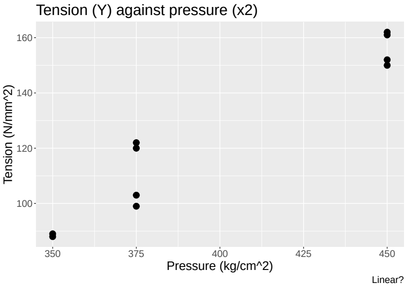
The latter two concepts will be emphasized when we talk about hypothesis tests later.

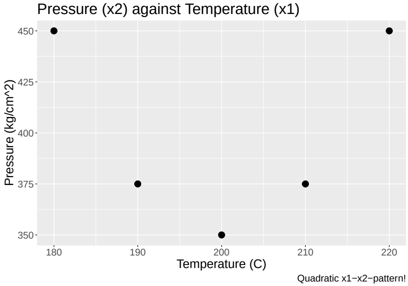
Example: Elasticity

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







The quadratic relationship between x_1 and x_2 makes the (marginal) relationship between Y and x_1 look quadratic as well!

See `lecture03_ex1_elasticity.html` for a rotatable 3D-plot.

- 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 $(*)$ is suitable for the **joint** relationship.

Elasticity: estimates

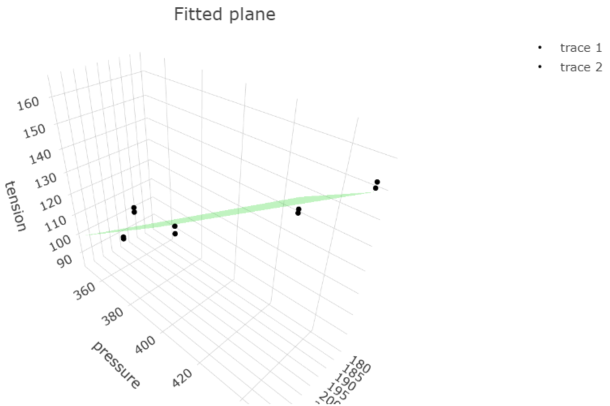
Y = tension (N/mm^2), x_1 = temperature (°C), x_2 = pressure (kg/cm^2).
Model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$, $i = 1, \ldots, 10$, $\epsilon_i \sim N(0, \sigma^2)$

Variable		est.	s.e.	95 % C.I.	unit
intercept	β_0	-215.7	31.9	$(-291.1, -140.2)$	N/mm^2
temperature	β_1	0.41	0.13	$(0.10, 0.72)$	$\frac{\text{N/mm}^2}{°\text{C}}$
pressure	β_2	0.65	0.04	$(0.54, 0.75)$	$\frac{\text{N/mm}^2}{\text{kg/cm}^2}$
resid.std.dev	σ	5.90	df $= 7$		N/mm^2

Fitted plane: $\hat{Y} = -215.7 + 0.41x_1 + 0.65x_2$.

Effect of temperature change

Increase the temperature by Δ_1 °C from x_{01} to $x_{01} + \Delta_1$
while keeping the pressure fixed at x_{02}:
$\hat{Y}_{\text{old}} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02}$
$\hat{Y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 (x_{01} + \Delta_1) + \hat{\beta}_2 x_{02}$
$\hat{Y}_{\text{new}} - \hat{Y}_{\text{old}} = \hat{\beta}_1 \Delta_1 = 0.41\Delta_1$ (N/mm^2) regardless of the pressure.


Fitted plane
trace 1
trace 2
tension
160
150
140
130
120
110
100
90
pressure
360
380
400
420
180
185
190
195

Predictions

If temperature = 200 °C and pressure = 400 kg/cm^2,
$\mathbf{x}_0 = [1 \quad 200 \quad 400]$

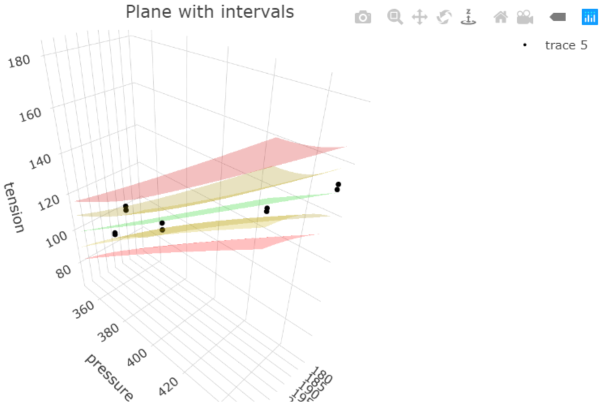
- what is the expected tension?
  $\hat{Y}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}} = -215.7 + 0.41 \cdot 200 + 0.65 \cdot 400 = 124.6$ (N/mm^2).
- What tension values might we observe?
  $\hat{Y}_{\text{pred}_0} = \mathbf{x}_0\hat{\boldsymbol{\beta}} + \epsilon_0 = 124.6 + \epsilon_0$ (N/mm^2).

	estimate	s.e.	95 % interval	
on average	$\hat{Y}_0 = 124.6$	1.87	$(120.2, 129.0)$	conf.
single obs.	$\hat{Y}_{\text{pred}_0} = 124.6 + \epsilon_0$	6.19	$(110.0, 139.2)$	pred.

Note: $6.19 = \sqrt{5.90^2 + 1.87^2}$.

Intervals for the plane

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



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 new variables, one less than the number of categories, e.g., x_{weight} is replaced by the two **dummy variables**

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

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

x_{weight}	x_{under}	x_{over}
normalweight	0	0
underweight	1	0
overweight	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:
$Y_i = \beta_0 + \beta_{\text{under}} x_{\text{i,under}} + \beta_{\text{over}} x_{\text{i,over}} + \epsilon_i$.

- "normalweight" 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 relative to the reference category.

$$E(Y \mid \text{normalweight}) = \beta_0$$
$$E(Y \mid \text{underweight}) = \beta_0 + \beta_{\text{under}}$$
$$E(Y \mid \text{overweight}) = \beta_0 + \beta_{\text{over}}$$

Warning

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.

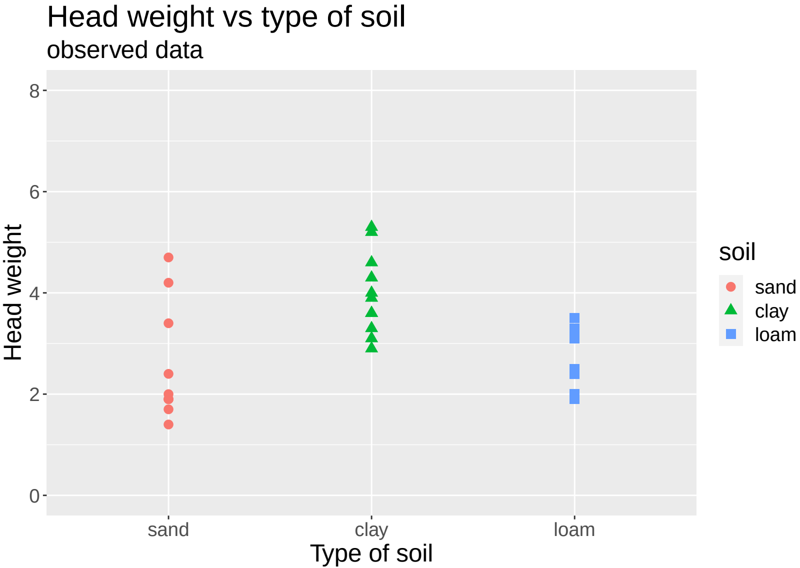
R calls categorical variables `factors`.
Categories for a categorical variable are called `levels`.

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			
			clay	60	5.2			

Head weight vs type of soil



Cabbage: soil model and estimates

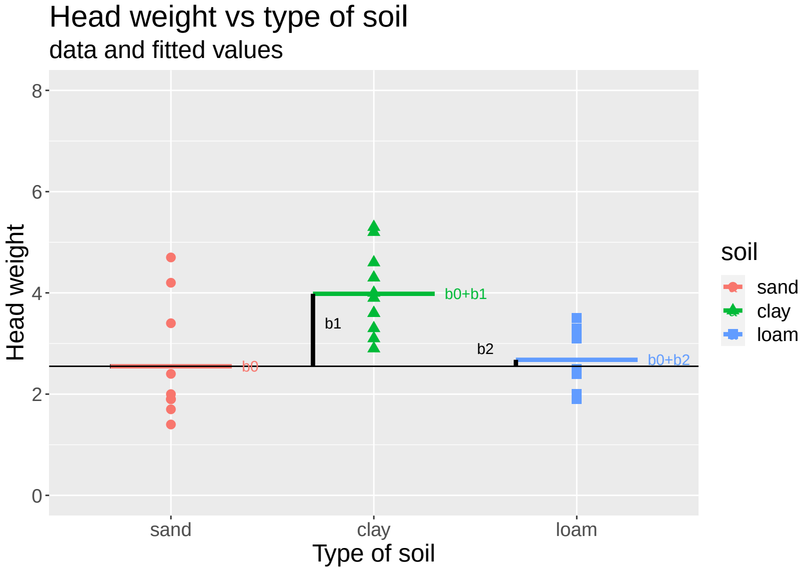
$Y =$ head weight, $x_1 = \begin{cases} 1 & \text{clay} \\ 0 & \text{not clay} \end{cases}$, $x_2 = \begin{cases} 1 & \text{loam} \\ 0 & \text{not loam} \end{cases}$

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	$(1.97, 3.13)$
clay (vs sand)	β_1	1.43	0.39	$(0.63, 2.24)$
loam (vs sand)	β_2	0.13	0.41	$(-0.72, 0.98)$
resid.std.dev	σ	0.90	df $= 27$	

Fitted "line":

$$\hat{Y} = 2.55 + 1.43x_1 + 0.13x_2 = \begin{cases} 2.55 & = 2.55, & \text{sand} \\ 2.55 + 1.43 & = 3.98, & \text{clay} \\ 2.55 + 0.13 & = 2.68, & \text{loam} \end{cases}$$


Head weight vs type of soil
data and fitted values
Head weight
8
6
4
2
0
b0
b1
b0+b1
b2
b0+b2
sand
clay
loam
Type of soil
soil
sand
clay
loam

Predictions

- What is the average head weight for the different soil types?

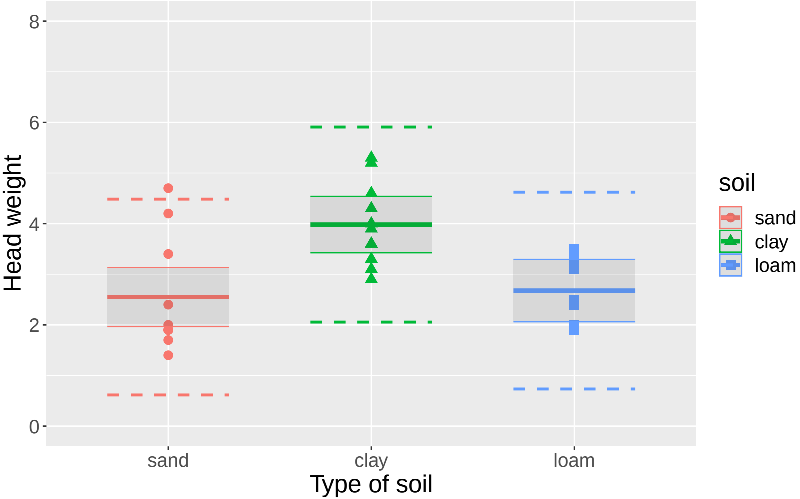
on average	estimate	s.e.	95 % C.I.
soil	$\hat{Y}_{\text{soil}} = \hat{\beta}_0 = 2.55$	0.28	$(1.97, 3.14)$
clay	$\hat{Y}_{\text{clay}} = \hat{\beta}_0 + \hat{\beta}_1 = 3.98$	0.27	$(3.43, 4.54)$
loam	$\hat{Y}_{\text{loam}} = \hat{\beta}_0 + \hat{\beta}_2 = 2.68$	0.30	$(2.06, 3.29)$

- What head weights might we observe?

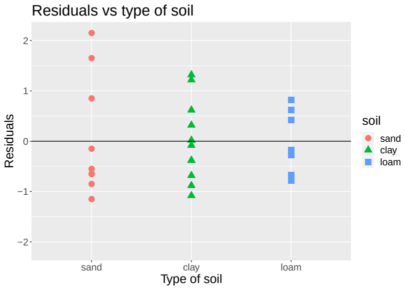
single head	estimate	s.e.	95 % P.I.
soil	$\hat{Y}_{\text{pred}_{\text{soil}}} = 2.55 + \epsilon_0$	0.94	$(0.62, 4.48)$
clay	$\hat{Y}_{\text{pred}_{\text{clay}}} = 3.98 + \epsilon_0$	0.94	$(2.06, 5.91)$
loam	$\hat{Y}_{\text{pred}_{\text{loam}}} = 2.68 + \epsilon_0$	0.95	$(0.73, 4.62)$

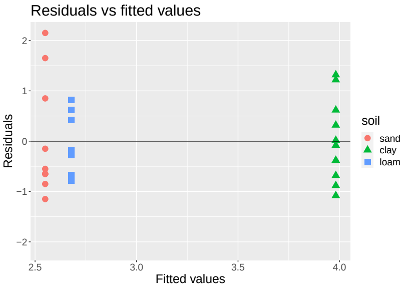
Head weight vs type of soil

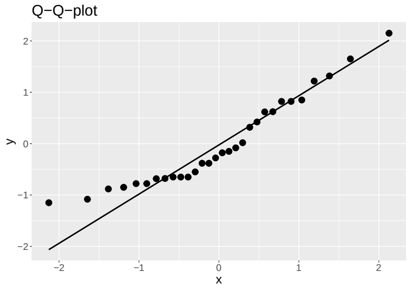
data, fitted line, confidence and prediction intervals

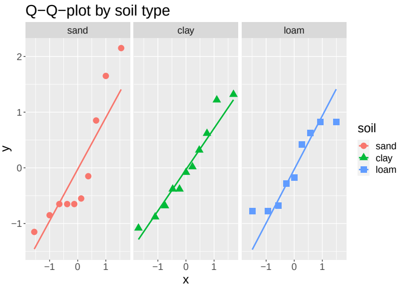


Basic residual analysis





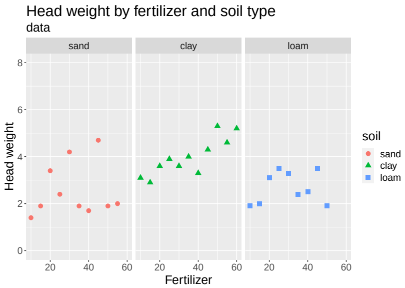


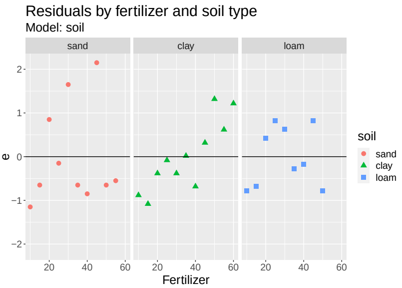


Conclusions

- The difference in average head weight between sand and loam is not necessarily different from 0 (the confidence interval $I_{\beta_2} = (-0.73, 0.98)$ contains 0).
- The residual variability in loam seems smaller than in the other two soil types.
- The residuals are not very un-normal.

What about fertilizer?





New questions

- There is a linear pattern in the residuals, at least for clay.
- Would adding the amount of fertilizer to the model improve the fit?

Cabbage: soil and fertilizer model

$$Y = \text{head weight},$$

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

$$x_3 = \text{fertilizer.}$$

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

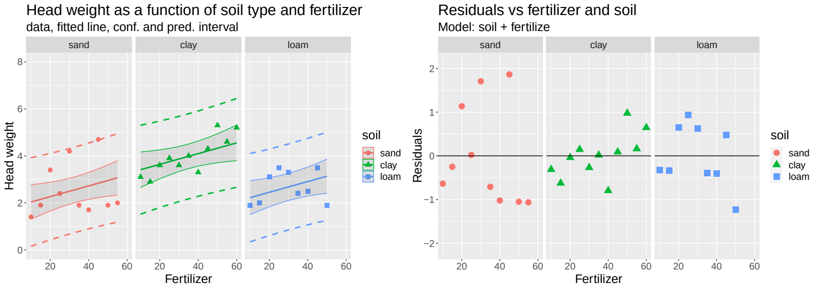
Variable	parameter	estimate	s.e.	95 % C.I.
intercept (sand)	β_0	1.81	0.44	$(0.91, 2.70)$
clay (vs sand)	β_1	1.37	0.37	$(0.61, 2.13)$
loam (vs sand)	β_2	0.18	0.39	$(-0.61, 0.98)$
fertilize	β_3	0.023	0.011	$(-0.001, 0.045)$
resid.std.dev	σ	0.84	df $= 26$	

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:

$$\hat{Y} = 1.81 + 1.37x_1 + 0.18x_2 + 0.023x_3$$

$$= \begin{cases} 1.81 + 0.023x_3 & \text{sand} \\ 1.81 + 1.37 + 0.023x_3 & \text{clay} \\ 1.81 + 0.18 + 0.023x_3 & \text{loam} \end{cases}$$



Even more questions

- There seems to still be a positive, but smaller, linear pattern in the residuals for clay.
- There is now a possible negative relationship in loam.
- Maybe the effect of the fertilizer is not the same for all three soil types.
- Solution: add interaction terms to the model.

Interaction

What if

- the effect of changing the temperature **depends on the pressure**?
- the effect of changing the amount of fertilizer **depends on the soil type**?

Interaction terms

content...

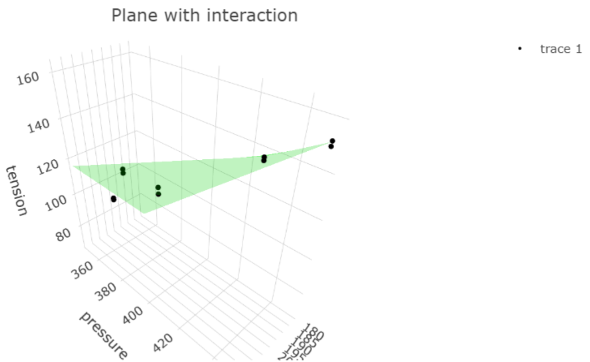
Example: Elasticity with interaction

Add the **interaction** term $x_3 = x_1 \cdot x_2 =$ temperature × pressure to the model.

$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$, $i = 1, \ldots, 10$, $\epsilon_i \sim N(0, \sigma^2)$

Variable		est.	s.e.	95 % C.I.
intercept	β_0	-1071.2	221.6	$(-1613.4, -529.0)$
temperature	β_1	4.69	1.11	$(1.98, 7.40)$
pressure	β_2	2.61	0.51	$(1.37, 3.86)$
temp×press	β_3	-0.0098	0.0025	$(-0.016, -00036)$
resid.std.dev	σ	3.40	df $= 6$	

Fitted plane: $\hat{Y} = -1071.2 + 4.69 x_1 + 2.61 x_2 - 0.0098 x_1 x_2$.


Plane with interaction
trace 1
tension
160
140
120
100
80
pressure
360
380
400
420

Effect of temperature change: interaction

Increase the temperature by Δ_1 °C from x_{01} to $x_{01} + \Delta_1$ while keeping the pressure fixed at x_{02}:

$$\begin{aligned}
\hat{Y}_{\text{old}} &= \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \hat{\beta}_3 x_{01} x_{02} \\
\hat{Y}_{\text{new}} &= \hat{\beta}_0 + \hat{\beta}_1 (x_{01} + \Delta_1) + \hat{\beta}_2 x_{02} + \hat{\beta}_3 (x_{01} + \Delta_1) x_{02} \\
\hat{Y}_{\text{new}} - \hat{Y}_{\text{old}} &= (\hat{\beta}_1 + \hat{\beta}_3 x_{02}) \Delta_1 \text{ depends on the pressure!}
\end{aligned}$$

Temperature effect for some different pressures:

$$\begin{aligned}
x_{02} = 350 &: \hat{Y}_{\text{new}} - \hat{Y}_{\text{old}} = (\hat{\beta}_1 + \hat{\beta}_3 \cdot 350)\Delta_1 = 1.24\Delta_1 \\
x_{02} = 400 &: \hat{Y}_{\text{new}} - \hat{Y}_{\text{old}} = (\hat{\beta}_1 + \hat{\beta}_3 \cdot 400)\Delta_1 = 0.75\Delta_1 \\
x_{02} = 450 &: \hat{Y}_{\text{new}} - \hat{Y}_{\text{old}} = (\hat{\beta}_1 + \hat{\beta}_3 \cdot 450)\Delta_1 = 0.26\Delta_1
\end{aligned}$$

Note: without the interaction we always had $0.41\Delta_1$.

Predictions: with interaction

If temperature = 200 °C and pressure = 400 kg/cm^2,
$\mathbf{x}_0 = [1 \quad 200 \quad 400 \quad 80000]$

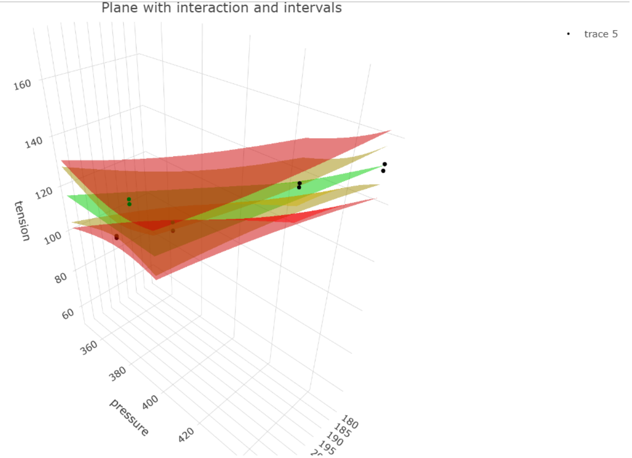
- what is the expected tension?
  $\hat{Y}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}} =$
  $= -1071.2 + 4.69 \cdot 200 + 2.61 \cdot 400 - 0.0098 \cdot 80000 = 124.6.$
- What tension values might we observe?
  $\hat{Y}_{\text{pred}_0} = \mathbf{x}_0\hat{\boldsymbol{\beta}} + \epsilon_0 = 124.6 + \epsilon_0.$

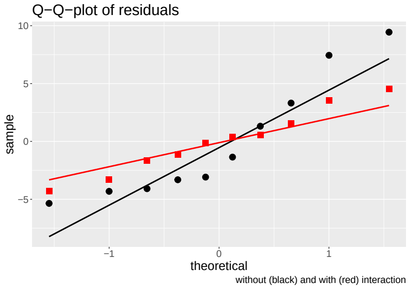
	estimate	s.e.	95 % interval	
on average	$\hat{Y}_0 = 124.6$	1.08	$(122.0, 127.2)$	conf.
single obs.	$\hat{Y}_{\text{pred}_0} = 124.6 + \epsilon_0$	3.57	$(115.9, 133.3)$	pred.

Note: $3.57 = \sqrt{3.41^2 + 1.08^2}$.

Note that both $\hat{\sigma}$ and all the intervals have become narrower. This model fits closer to the data!

Too close? Overfitting? Next week...


Plane with interaction and intervals
trace 5
tension
160
140
120
100
80
60
pressure
360
380
400
420
180
185
190
195



Conculsions

The model with interaction has

- a β_3 for the interaction term that is not zero,
- smaller residuals,
- residuals closer to a normal distribution

and is thus, probably, better than the model without interaction.
But see Lecture 4, 5 and 6 before the final conclusion...

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.

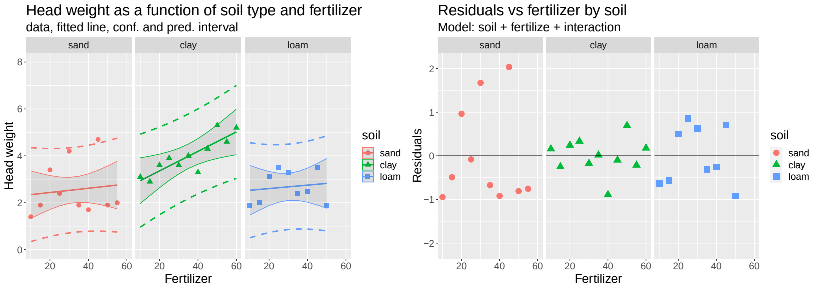
Cabbage: soil and fertilizer interaction: estimates

Variable	parameter	estimate	s.e.	95 % C.I.
intercept (sand)	β_0	2.25	0.65	$(0.90, 3.60)$
clay	β_1	0.27	0.90	$(-1.58, 2.12)$
loam	β_2	0.20	0.96	$(-1.78, 2.19)$
fertilize (sand)	β_3	0.009	0.029	$(-0.027, 0.047)$
fert:clay	β_4	0.033	0.024	$(-0.018, 0.083)$
fert:loam	β_5	-0.002	0.028	$(-0.060, 0.057)$
resid.std.dev	σ	0.84	df $= 24$	

Fitted lines:

$$\hat{Y} = 2.25 + 0.37x_1 + 0.20x_2 + 0.009x_3 + 0.033x_1x_3 - 0.002x_2x_3$$

$$= \begin{cases} 2.25 + 0.009x_3 & \text{sand} \\ 2.25 + 0.37 + (0.009 + 0.033)x_3 & \text{clay} \\ 2.25 + 0.20 + (0.009 - 0.002)x_3 & \text{loam} \end{cases}$$



Some conclusions from the residual analysis

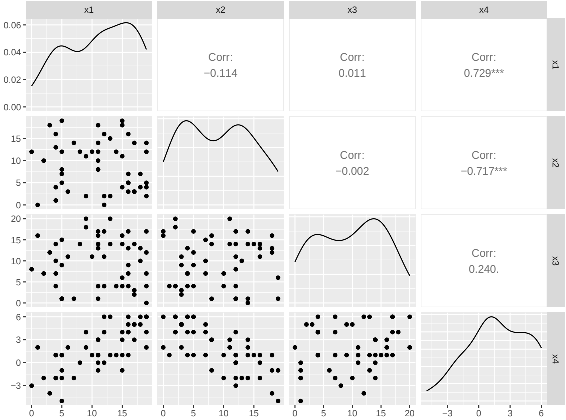
- The residual variability in clay is now as small as that for loam! The fertilizer explained most of it.
- The residual variability in sand is still large and un-explained.
- The residuals are slightly closer to a normal distribution.

Collinearity among x-variables

- The matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ must be invertible.
- If $\mathbf{X}^\mathsf{T}\mathbf{X}$ is singular there is no unique solution. If it is nearly singular the solution is unstable.
- Is a problem when any of the x-variables is (almost) a linear combination of some of the other x-variables.
- β-estimates will then not exist or will have huge variance.
- Correlated x-variables "compete" (one variable might be necessary if the other is not in the model, but not if both are in the model, etc.)
- Found by:
  - pairwise plots and correlations of all (potential) x-variables against each other,
  - Variance Inflation Factors (VIF or GVIF)
- Solution: Remove the most problematic variables

Pairwise dependant x-variables

- plot all pairs of x-variables,
- calculate all pairwise correlations between numerical x-variables



- Rule of thumb: correlations $|\rho| > 0.7$ might be worrying.
- $\rho(X_1, X_4) = 0.729 > 0.7$
- $\rho(X_2, X_4) = -0.717 < -0.7$

Linear combinations between several x-variables

- Difficult to find with pairwise analysis.
- For each x-variable, $x_1, \ldots, x_p$, fit a linear model with the column $\mathbf{X}_{.j}$ as dependent variable and all the others as explanatory variables, giving the predicted values $\hat{\mathbf{X}}_{.j}$.
- The correlation $R_j = \rho(\mathbf{X}_{.j}, \hat{\mathbf{X}}_{.j})$ should preferrably be close to zero. If it is large we have a problem.
- The square of the correlation, $0 \leq R_j^2 \leq 1$, is the amount of variability in covariate x_j that can be explained by the other x-variables.

  C.f. the rule of thumb, $\rho(X_j, X_k)^2 > 0.7^2 = 0.49 \approx 50\,\%$.

Variance Inflation Factor, VIF

The Variance Inflation Factor (VIF) for covariate x_j is defined as

$$\mathsf{VIF}_j = \frac{1}{1 - R_j^2}$$

- $1 \leq \mathsf{VIF}_j$ should ideally be close to one.
- $\sqrt{\mathsf{VIF}_j}$ indicates how many times larger the standard error $d(\hat{\beta}_j)$ is because of the dependance with the other x-variables.
- Rules of thumb: 2 ($R_j^2 = 50\,\%$), 5 ($80\,\%$), 10 ($90\,\%$)

Variable	VIF		
x_1	58.8	large	
x_2	56.8	large	
x_3	8.3		
x_4	135.4	very large	Do not use x_4 in a model with x_1, x_2, x_3.

Generalised Variance Inflation Factor, GVIF

- When we have categorical x-variables there is a structural dependancy between the resulting dummy variables.
- The same is true when we have added interaction terms to the model.
- Calculate GVIF using this imposed structure of the data.
- We get GVIF with degrees of freedom f as the number of dummy variables involved in the factor variable and/or the interaction terms.
- Rules of thumb for GVIF$^{1/2f}$: $\sqrt{2}$, $\sqrt{5}$, $\sqrt{10}$

Variable	GVIF	f	GVIF$^{1/2f}$	
x_1 : soil	1.02	2	1.005	Small
x_2 : fertilize	1.01	1	1.01	Small