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

Lecture 1, spring 2023

Linear regression: assumptions and estimates

https://canvas.education.lu.se/courses/23000

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

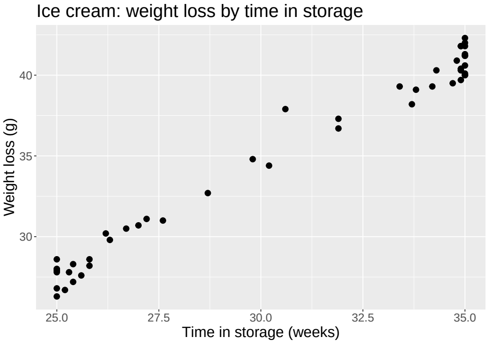
20/3-23

Simple Linear Regression

We measure two variables, x and Y. How does the value of Y depend on the value of x? Is there a linear relationship? How can we estimate this relationship using observed data?

Example: Ice cream

An ice cream manufacturer suspects that storing ice cream at low temperatures leads to weight loss.

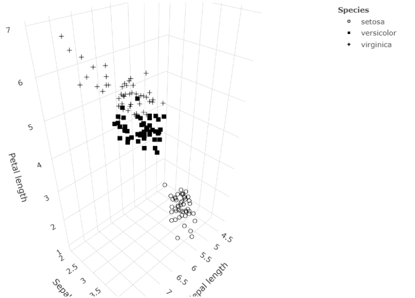


Multiple linear regression

We measure several variables, $x_1, \ldots, x_p$, and Y. How does the value of Y depend on the values of $x_1, \ldots, x_p$?

Example: Iris

The petal length depends on sepal length, sepal width and species.



Basic assumptions

- Y: continuous dependent variable, "response" or "outcome", assumed random.
- $x_1, \ldots, x_p$: explanatory variables, "covariates"; assumed non-random.
- We *hypothesize* that Y has a linear relationship with $x_1, \ldots, x_p$, on average, and follows the linear model:

$$E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

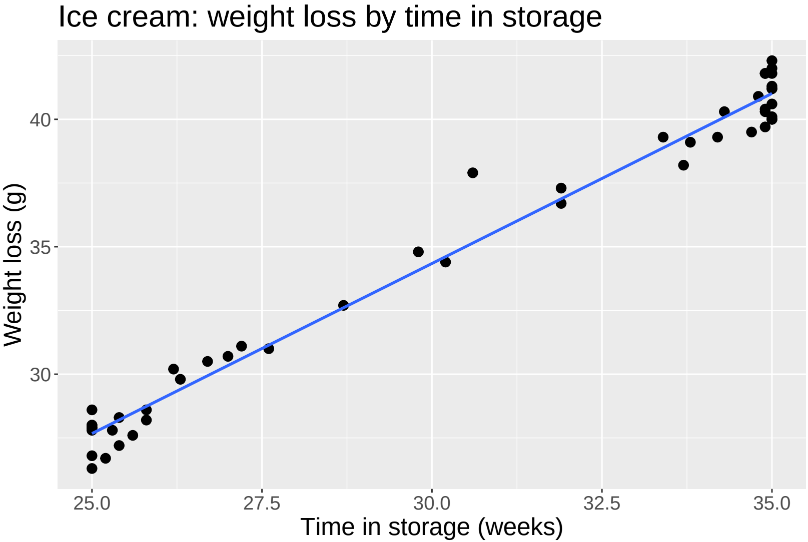
- $\beta_0, \beta_1, \ldots, \beta_p$: unknown parameters, assumed non-random.
- $\beta_0 =$ intercept; $E(Y)$ when $x_1 = \cdots = x_p = 0$,
- $\beta_1, \ldots, \beta_p =$ slopes in the corresponding x-directions; the additive change in $E(Y)$ when the corresponding x-variable is increased by 1 unit and the others are held fixed.

Model: multiple linear regression

- We denote with Y_i, where $i = 1, \ldots, n$, the ith observation from a set of n measurements of Y.
- We denode with x_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$, the corresponding ith observation of the jth x-variable.
- The model for a generic observation Y_i is

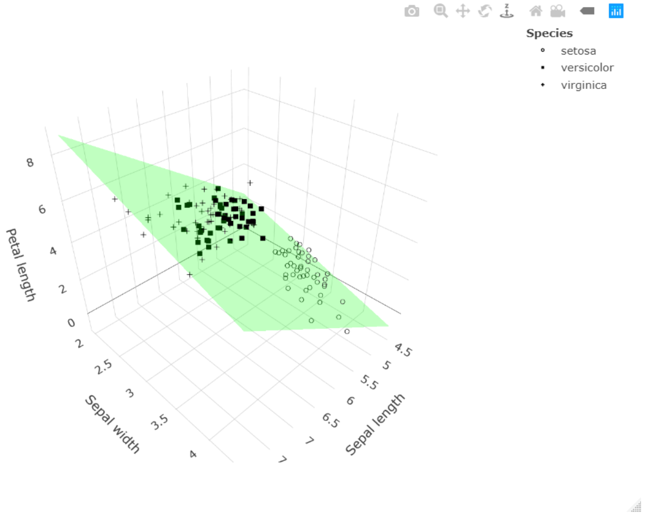
$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, \ldots, n$$

where ϵ_i is the “measurement error” which contains all the random variation not explained by the linear model.



data and fitted line

Iris: petal length as function of sepal length and width, $p = 2$



Assumptions for the measurement error

Besides linearity, we also assume for all $i = 1, ..., n$

$$E(\epsilon_i) = 0 \qquad E(Y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \mu_i$$

$$V(\epsilon_i) = \sigma^2 \qquad V(Y_i) = \sigma^2$$

$\epsilon_i \sim N(0, \sigma^2)$ $\qquad$ $Y_i \sim N(\mu_i, \sigma^2)$

ϵ_i are pairwise independent $\qquad$ Y_i are pairwise independent

A note on notation

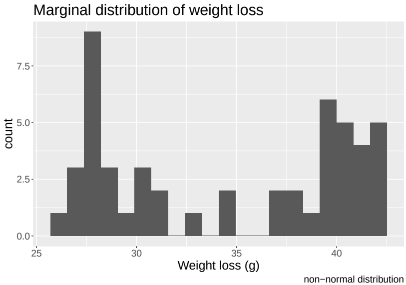
Strictly speaking, we assume properties of Y_i *conditional* on $X_1 = x_{i1}$:

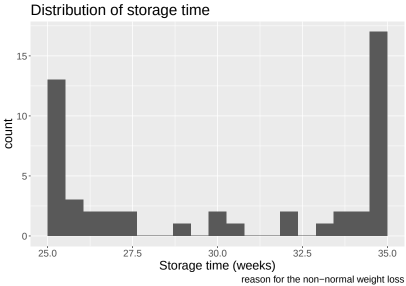
$$E(Y_i \mid X_1 = x_{i1}, \ldots, X_p = x_{ip}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \mu_i,$$
$$V(Y_i \mid X_1 = x_{i1}, \ldots, X_p = x_{ip}) = \sigma^2,$$
$$Y_i \mid X_1 = x_{i1}, \ldots, X_p = x_{ip}) \sim N(\mu_i, \sigma^2)$$

We consider this notation as implicit and always write $E(Y_i)$, etc., in place of $E(Y_i \mid X_1 = x_{i1}, \ldots)$ . . . except on the next slide. . .

Warning!

Our assumptions imply that the conditional distribution of $Y_i \mid X = x_i$ is Normal but we don't know the marginal distribution of Y_i ignoring X!





The marginal distribution of Y is clearly not Normal but this is just due to the strange distribution of the x-values.

To assess normality you should instead inspect residuals (introduced later), which means you cannot always see that your model will be wrong (or right!) until after you have fitted it!

Multiple linear regression with matrices

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

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

$$\sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

(n-dimensional multivariate normal distribution)

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

the covariance matrix $\mathrm{Var}(\boldsymbol{\epsilon})$ is given by

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

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

Least squares estimates: simple linear without matrices

Find β_0, β_1 that minimize the loss function

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n}(Y_i - E(Y_i))^2 = \sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 x_i))^2.$$

Partial derivatives

Find the minimum by solving the linear equation system

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i) = 0 \Leftrightarrow \qquad n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}x_i(Y_i - \beta_0 - \beta_1 x_i) = 0 \Leftrightarrow \quad \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Least squares estimates: with matrices

Find $\boldsymbol{\beta}$ that minimizes $Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$.

Partial derivatives

Expand $Q(\boldsymbol{\beta})$ and use the fact that all the terms are scalar

$$
\begin{aligned}
Q(\boldsymbol{\beta}) &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\
\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} &= -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0}
\end{aligned}
$$

The solution satisfies the normal equations: $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$.

$$
\begin{pmatrix}
n & \sum_{i=1}^n x_{i1} & \cdots & \sum_{i=1}^n x_{ip} \\
\sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \cdots & \sum_{i=1}^n x_{i1}x_{ip} \\
\vdots & \vdots & \vdots & \vdots \\
\sum_{i=1}^n x_{ip} & \sum_{i=1}^n x_{i1}x_{ip} & \cdots & \sum_{i=1}^n x_{ip}^2
\end{pmatrix}
\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}
=
\begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_{i1}Y_i \\ \vdots \\ \sum_{1=1}^n x_{ip}Y_i \end{pmatrix}
$$

- **Parameter estimates**: $\hat{\boldsymbol{\beta}}$ is the solution to the normal equations:

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

- **Predicted values** = fitted line (plane):

$$\hat{Y}_i = \hat{E}(Y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip} \qquad \hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

- **Residuals**: the difference between observations and predictions:

$$e_i = Y_i - \hat{Y}_i \qquad \mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

- **Residual variance**: s^2 is an estimate of the variance of the error, a measure of the "residual variability" unexplained by the model.

$$\hat{\sigma}^2 = s^2 = \frac{Q(\hat{\boldsymbol{\beta}})}{n-(p+1)} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-(p+1)} = \frac{\mathbf{e}'\mathbf{e}}{n-(p+1)}$$

Note: $p+1$ is the total number of β-parameters in the model.

Properties of parameter estimates

Note: For a constant matrix $\mathbf{A}$ and a random matrix $\mathbf{Y}$ we have $E(\mathbf{AY}) = \mathbf{A}E(\mathbf{Y})$ and $\mathrm{Var}(\mathbf{AY}) = \mathbf{A}\mathrm{Var}(\mathbf{Y})\mathbf{A}'$.

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

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Var}(\mathbf{Y})\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \sigma^2\mathbf{I} \cdot \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\
&= \begin{pmatrix} V(\hat{\beta}_0) & \cdots & C(\hat{\beta}_0, \hat{\beta}_p) \\ \vdots & \ddots & \vdots \\ C(\hat{\beta}_p, \hat{\beta}_0) & \cdots & V(\hat{\beta}_p) \end{pmatrix}
\end{aligned}$$

When $p = 1$:

$$V(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right), \quad V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$C(\hat{\beta}_0, \hat{\beta}_1) = C(\hat{\beta}_1, \hat{\beta}_0) = -\bar{x}V(\hat{\beta}_1)$$

For a specific set of x-values, $\mathbf{x}_0 = \begin{pmatrix} 1 & x_{01} & \ldots & x_{0p} \end{pmatrix}$, we have

- on average (the fitted line) $\hat{Y}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}}$:

$$E(\hat{Y}_0) = \mathbf{x}_0 E(\hat{\boldsymbol{\beta}}) = \mathbf{x}_0\boldsymbol{\beta} = \beta_0 + \beta_1 x_{01} + \ldots + \beta_p x_{0p} = \mu_0,$$

$$V(\hat{Y}_0) = \mathbf{x}_0 \mathrm{Var}(\hat{\boldsymbol{\beta}})\mathbf{x}_0' = \sigma^2 \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'$$

$$[p=1] = \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)$$

- for a new observation $\hat{Y}_{\mathsf{pred}_0} = \mathbf{x}_0\hat{\boldsymbol{\beta}} + \epsilon_0$:

$$E(\hat{Y}_{\mathsf{pred}_0}) = \mathbf{x}_0 E(\hat{\boldsymbol{\beta}}) + E(\epsilon_0) = \mathbf{x}_0\boldsymbol{\beta} = \mu_0,$$

$$V(\hat{Y}_{\mathsf{pred}_0}) = \mathbf{x}_0 \mathrm{Var}(\hat{\boldsymbol{\beta}})\mathbf{x}_0' + V(\epsilon_0) = \sigma^2(1 + \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0')$$

$$[p=1] = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)$$

Standard error

The standard error (s.e. = medelfel), $d(\hat{\theta}) = \sqrt{\hat{V}(\hat{\theta})}$, of an estimate $\hat{\theta}$, replaces any unknown parameters in $V(\hat{\theta})$ by their estimates. Here, replace σ by $\hat{\sigma} = s$.

Comments

- $\hat{\boldsymbol{\beta}}$ exists only when $\mathbf{X}'\mathbf{X}$ is non-singular. Near-singularity makes all β-estimates very uncertain.
- Never calculate $(\mathbf{X}'\mathbf{X})^{-1}$ "by hand". Take a course in Numerical analysis (or rely on R).
- The uncertainty of the $\boldsymbol{\beta}$-estimates is mostly due to the sample size, n, and the structure of the x-variables, as expressed in $(\mathbf{X}'\mathbf{X})^{-1}$.
- The uncertainty of the predictions $\hat{Y}_0$ is also due to how far $\mathbf{x}_0$ is from the midpoint of the x-variables.
- $V(\hat{Y}_{\text{pred}_0}) \to \sigma^2 > 0$ when $V(\hat{Y}_0) \to 0$.

Ice cream: estimates

Y = weight loss (g), x = storage time (weeks).

Model $Y = \beta_0 + \beta_1 x + \epsilon$

Variable	parameter	estimate	s.e.	unit
intercept (time = 0)	β_0	-5.7	0.81	g
storage time	β_1	1.33	0.03	g/week
resid.std.dev	σ	0.80		g

Fitted line: $\hat{Y} = -5.7 + 1.33x$.

Predictions

If we store the ice cream for $x_0 = 34$ weeks, how much weight loss can we expect on average? in a single package?

	estimate	s.e.	unit
on average	$\hat{Y}_0 = -5.7 + 1.33 \cdot 34 = 39.7$	0.15	g
single package	$\hat{Y}_{\text{pred}_0} = 39.7 + \epsilon_0$	0.82	g

Note: $0.82 = \sqrt{0.80^2 + 0.15^2}$.