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

Lecture 1, spring 2019

Simple linear regression, parameter estimates, properties

www.maths.lth.se/matstat/kurser/masm22/

www.maths.lth.se/matstat/kurser/fmsn40/

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

25/3-19

Formalia

Teachers

Lecturer: Anna Lindgren, MH:136, anna.lindgren@matstat.lu.se
Assistants: Vasilii Goriachkin, Amanda Nilsson

Suggested literature (freely available as e-books)

- Rawlings, Pantula, Dickey: Applied Regression Analysis - A Research Tool, 2ed 1998, Springer
- Agresti, A. An Introduction To Categorical Data Analysis, 2ed Wiley, 2007

Course registration

- Course register yourself at `www.student.lu.se` or `www.student.ladok.se`.
- or check yourself off on the circulating list and we will course register you.
- or we will course register you if you do the first lab.

Cannot be registered?

- You must have passed a basic course in statistics or mathematical statistics in order to take this course. Contact `studierektor@matstat.lu.se`
- Forgot to apply for the course?
  - LTH engineering students (FMSN30/FMSN40): contact your program so they can register you.
  - Everybody else, incl. PhD students, (MASM22): make a late application at `www.antagning.se` no later than 31 March.

Course details

- Lectures
- Three mandatory computer exercises, check the schedule
- Project help (not-compulsory)
- Three projects (groups of **two** students)
  - Project 1: Linear regression. Written report with peer assessment 15 April. Final version due 17 April.
  - Project 2: Logistic regression: Written report with peer assessment 15 May. Final version 17 May.
  - Project 3: Some extended model:
    MASM22/FMSN30: oral presentation ca 15 min 27, 28 and 29 May
    FMSN40: project plan due 17 April, data gathering done 21 May, oral presentation ca 20 min 27, 28 and 29 May.
- Individual oral exam 3 – 20 June (Midsummer).

Sign-up for labs, project 3 presentations and oral exam

- `https://matstat.sam.cs.lth.se/Labs`
- MASM22/FMSN30: Sign-up for lab 1, 2 and 3 is open.
  FMSN40: you only have one lab group so no lab sign-up.
- Sign-up for the oral presentations for project 3 at 27–29 May and the oral exams in June will open after Easter.

You cannot change these bookings yourself. Contact Anna.

Handing in project reports

Follow the instructions given in the projects. Specifically:

- mail your project report to the correct adress:
  MASM22 and FMSN30: use `fmsn30@matstat.lu.se`,
  FMSN40: use `fmsn40@matstat.lu.se`,
- the subject *must* follow the instructions, including the exact spelling. We use the subject to automatically identify the course object (Projekt 1 $\neq$ Project1), create a group of the students involved and associate the report with them,
- Only send pdf and R files, not, e.g., Word, zip or png.

Mail not related to the project reports or sent to the wrong course will end up among the spam.

Why Modelling?

This course introduces some ideas to deal with modelling of dependencies between several variables.

- But why should we bother about modelling?
- After all:
  "All models are wrong, but some are useful." [George Box]
- Everything in Nature and in Society varies. Most of such variability cannot be captured with deterministic mathematics and physics.
- Statistics is a powerful mathematical science to extract information, make predictions, find relations in large amounts of data and to model knowledge about uncertain phenomena. It's the Mathematics of Uncertainty!
- On TED-talk someone even proposed to teach Statistics before Calculus![1]

[1] http://tinyurl.com/nw8uyo

Why Modelling?

- New Statistical journals are created each year.
- Dozens of new articles are published every day.
- Larger and increasingly complex models are created in order to deal with the increasing amount of data we are all exposed to.
- In this course we start with basic but still very relevant statistical models.
- Understanding the main tools for **linear models** is fundamental as these are also used for **nonlinear models**, with some technical modifications.
- My hope for this course is to serve as a useful basis to enlarge your understanding of **data dependencies**, check the importance of **statistical assumptions** and motivate you in studying more ambitious methods beyond the scope of this course.

Which models?

We will consider modelling the linear relationship between some **continuous** variable Y depending on:

1. a single variable X (simple linear regression);
2. several variables $X_1, ..., X_p$ (multiple linear regression);

We will also consider modelling the nonlinear relationship between some **binary** variable Y

- depending on $X_1, ..., X_p$ (logistic regression).

Additional models will be considered for **discrete** Y.
All the above will be complemented with specific **statistical inference** tools.

Long term goals

During the next months we learn how to answer the following:

- are my modelling/probabilistic assumptions satisfied?
- how do I test whether there is an **effect** of a variable on another variable?
- or in other words: how do I assess the statistical significance of said effect?
- is my model "explaining" enough of the phenomenon variability?
- is my model satisfactory or is it too big/small to represent variability?

To consider the above we introduce concepts and constructs that will help you study further modelling tools **beyond what is covered in this course**.

Data variables

Typical variables:

- **Continuous**: e.g. blood pressure, height, result of a physical measurement,...
- **Count**: discrete, counting the number of events, e.g. number of accidents,...
- **Categorical**: discrete or qualitative. Such as *gender (M/F)*, *political preference (left/right)*, *eye color (green/blue/brown/...)*
  - **nominal categorical** have no intrinsic ordering: *gender*, *eye colour*...
  - **ordinal categorical** have some natural ordering/ranking: *studies degree (bachelor/master/PhD)*, *satisfaction (not at all/sometimes/often)*.

We start with *simple linear regression*.

We gradually evolve to:

- multiple regression
- logistic regression
- Poisson and negative binomial regression
- Generalized linear models: GLMs contain all the above as special cases
- quantile regression

Simple linear regression models are indeed very simple.

However, this way we can easily construct tools that will be trivially extended to more complex models.

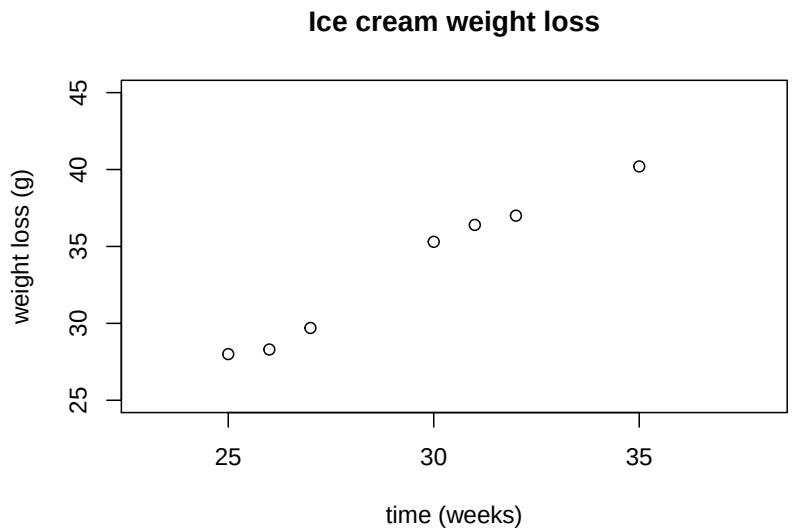
Simple Regression

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:

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

Time (weeks)	26	32	35	27	25	31	30
Weight loss (g)	28.3	37.0	40.2	29.7	28.0	36.4	35.3



Basic assumptions

Y = **continuous** dependent variable, "response" or "outcome", assumed **random**

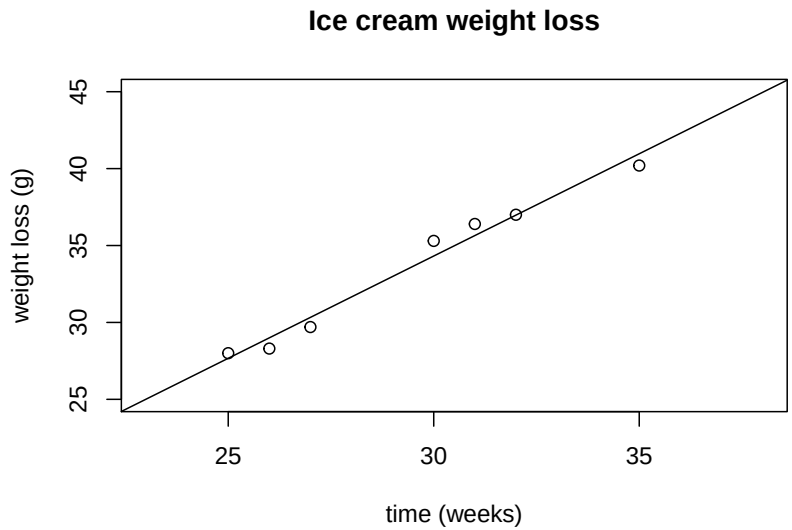
X = independent variable(s); assumed **non random**

We *hypothesize* that Y has a linear relationship with X, **on average**

That is we write a **Linear Model** $E(Y) = \beta_0 + \beta_1 X$

(β_0, β_1) = unknown parameters, assumed **non random**

β_0 = intercept; β_1 = slope.



- Notice if linearity between Y and X does not hold, worry not! You *might* still be able to transform Y and/or X so that the relationship becomes linear: e.g.

$$E(Y) = \beta_0 + \beta_1 \ln X$$

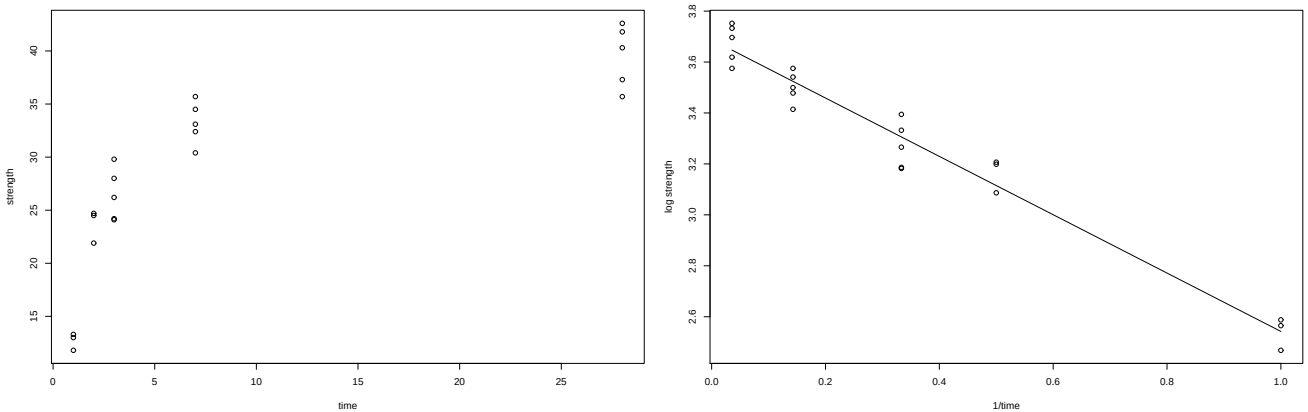
This is still a linear model, because linear in the parameters β_0, β_1.

Notation

In books you often find $E(Y \mid X) = \beta_0 + \beta_1 X$
This reads as "the expected value of Y, conditional on X, is linear".
We consider this notation as implicit and always write $E(Y)$ in place of $E(Y \mid X)$.

Example of linearization

ordinate (y-axis): cement tensile strength (kg/cm^2)
abscissa (x-axis): "curing time" (days): time needed for cement hardening.



(a) untransformed (b) transformed

Now take **ordinate**: ln(*strength*); **abscissa**: 1/time.
We are then perfectly allowed to fit:

$$E(\ln \text{strength}) = \beta_0 + \beta_1 \cdot 1/\text{time}$$

$$\widehat{\text{strength}} = e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot 1/\text{time}}$$

Assumptions for the measurement error

We denote with Y_i the ith observations from a set of n measurements of Y, $\{Y_1, Y_2..., Y_n\}$ and similarly for X_i.
A model for some measurement Y (not its expected response!):

$$Y = \beta_0 + \beta_1 X + \epsilon, \qquad \epsilon \text{ is "measurement error"}$$

that is for a generic observation i:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, ..., n$$

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

- $E(\epsilon_i) = 0$;
- $V(\epsilon_i) = \sigma^2$; where ϵ_i are pairwise independent.
- $\epsilon_i \sim N(0, \sigma^2)$

The assumptions imply that

- $E(Y_i \mid X_i) = E(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_i + E(\epsilon_i) = \beta_0 + \beta_1 X_i$
  because parameters and X_i are non-random.
- $V(Y_i \mid X_i) = V(\beta_0 + \beta_1 X_i + \epsilon_i) = 0 + V(\epsilon_i) = \sigma^2$
  because parameters and X_i are non-random.
- $Y_i \mid X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$
  because a normal distribution plus a constant is also a normal distribution.
- The Y_i are all (conditionally) independent
  because parameters and X_i are non-random while ϵ_i are independent.

Warning!

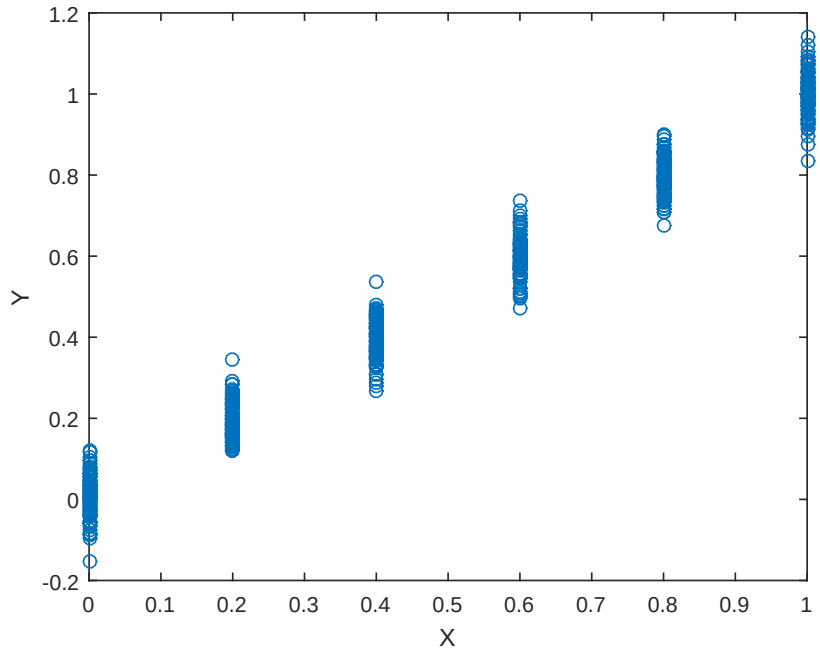
Our assumptions imply that $Y_i \mid X_i$ is Gaussian but we don't know the distribution of Y_i *marginally to* X_i.

That is to say we do not know the (marginal) distribution of Y_i.

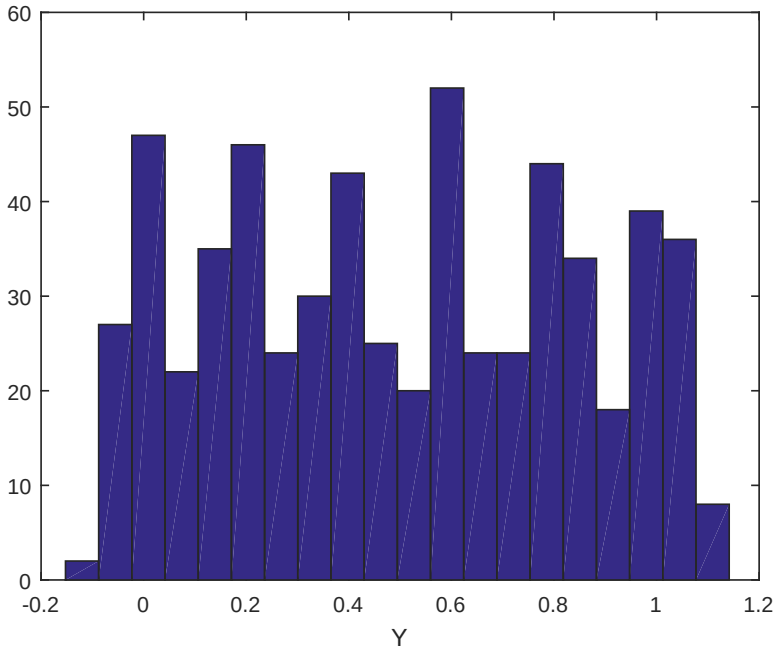
Simulation example (try yourself): set $\beta_0 = 0$ and $\beta_1 = 1$. Simulate $n = 1100$ draws

$$y_i \mid x_i \sim N(x_i, 0.05^2) \qquad i = 1, 2, ..., 1100$$

with x_i taking value 0 in one-hundred cases, value 0.2 in one-hundred cases, value 0.4 in one hundred cases etc.



Everything is linear, so all good...



The marginal distribution of Y. Clearly not Gaussian.

Warning!

This means that by plotting an histogram of the response Y you should not expect to necessarily obtain a Gaussian distribution, even when ϵ is Gaussian.

That's because plotting `hist(Y)` returns the marginal of Y.

To assess normality you should instead inspect **residuals** (introduced later).

This means you cannot always see that your model will be wrong until you have fitted it!

Least squares

Find β_0 and β_1 that minimize $\sum_{i=1}^{n}(Y_i - E(Y_i))^2$

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

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

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

Rearranging obtaines the **Normal equations**:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

$$\beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

Parameter estimates

Take as **estimates** the solutions to the normal equations:

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

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

Note that

$$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n (X_i - \bar{X}) Y_i - \sum_{i=1}^n (X_i - \bar{X}) \bar{Y}$$

$$= \sum_{i=1}^n (X_i - \bar{X}) Y_i - \bar{Y}(\sum_{i=1}^n X_i - n\bar{X}) = \sum_{i=1}^n (X_i - \bar{X}) Y_i$$

and $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \bar{X}) X_i$.

Predicted values

$$\hat{Y}_i = \hat{E}(Y_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

(for the weight-loss example: $\hat{\beta}_0 = -5.57$, $\hat{\beta}_1 = 1.33$)
...so for $X = 34 \rightarrow \hat{Y} = -5.57 + 1.33 \times 34 = 39.65$.

Residuals

$$e_i = Y_i - \hat{Y}_i = \text{observed} - \text{predicted}$$

Residual variance

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

($s^2 = 0.622$ and $s = 0.789$ in the numerical example)
Therefore s is an estimate of the standard deviation of the error, a measure of the "residual variability" unexplained by the model.

With R...

`lm()` is the main function to fit a linear model.

```
myfit <- lm(y ~ x)  # this is a comment...
summary(myfit) # access a rich output for myfit
```

See the commented R-file for example code and attend the labs to learn more!

A gentle reminder...

Table: All quantities are scalar. Let W and Z be random variables. a and b are constants.

$E(W \pm Z) = E(W) \pm E(Z)$ (linearity)
$E(a) = a$
$E(cW) = cE(W)$
$Var(W) = E(W^2) - (E(W))^2 = E(W - E(W))^2$
$Var(aW \pm b) = a^2 Var(W)$
$\mathrm{Var}(aW \pm bZ) = a^2 \mathrm{Var}(W) + b^2 Var(Z) \pm 2abCov(W, Z)$
$Cov(W, Z) = E[(W - E(W))(Z - E(Z))] = E(WZ) - E(W)E(Z)$

Properties of estimators

Unbiasedness

$$
\begin{aligned}
E(\hat{\beta}_1) &= \frac{\sum_{i=1}^n (X_i - \bar{X})(E(Y_i) - E(\bar{Y}))}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \bar{X})(\beta_0 + \beta_1 X_i - \beta_0 - \beta_1 \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_1 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} && = \beta_1 \\
E(\hat{\beta}_0) &= E(\bar{Y}) - \bar{X} E(\hat{\beta}_1) = \beta_0 + \beta_1 \bar{X} - \bar{X}\beta_1 && = \beta_0
\end{aligned}
$$

[We used the fact that the expectation of a constant is the constant itself; expectation is a linear operator.]

Variances

$$
\begin{aligned}
V(\hat{\beta}_1) &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2 V(Y_i)}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} = \sigma^2 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
V(\hat{\beta}_0) &= V(\bar{Y}) + \bar{X}^2 V(\hat{\beta}_1) - 2\bar{X} C(\bar{Y}, \hat{\beta}_1) \\
&= \frac{\sigma^2}{n} + \frac{\bar{X}^2 \sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} + 0 \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)
\end{aligned}
$$

In the previous slide we used the fact that $C(\bar{Y}, \hat{\beta}_1) = 0$, this is proved here,

Covariance

$$C(\bar{Y}, \hat{\beta}_1) = C\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \frac{\sum_{j=1}^{n}(X_j - \bar{X})Y_j}{\sum_{k=1}^{n}(X_k - \bar{X})^2}\right)$$

$$= \frac{1}{n} \cdot \frac{1}{\sum_{k=1}^{n}(X_k - \bar{X})^2} \cdot \sum_{i=1}^{n}\sum_{j=1}^{n}(X_j - \bar{X}) \cdot \underbrace{C(Y_i, Y_j)}_{V(Y_i) \text{ for } i=j;0 \text{ for } i\neq j}$$

$$= \frac{1}{n} \cdot \frac{1}{\sum_{k=1}^{n}(X_k - \bar{X})^2} \cdot \underbrace{\sum_{i=1}^{n}(X_i - \bar{X})V(Y_i)}_{=\sigma^2(\sum X_i - n\bar{X})=0} = 0$$

Distributions

- Important property of a Normal distribution: any linear combination of independent normal variables is normally distributed.
- Because the β's are linear combinations of the Y_i's (which are assumed normal) we have the distributions below.

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right))$$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2})$$

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0 \sim N(\beta_0 + \beta_1 X_0, \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right))$$

$$\begin{aligned} \hat{Y}_{\text{pred}_0} &= \hat{\beta}_0 + \hat{\beta}_1 X_0 + \epsilon_0 \\ &\sim N(\beta_0 + \beta_1 X_0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)) \end{aligned}$$