SF2930 - Regression analysis

KTH Royal Institute of Technology, Stockholm

Lecture 1 – Introduction to the course and to regression analysis (MPV 2)

March 10, 2022

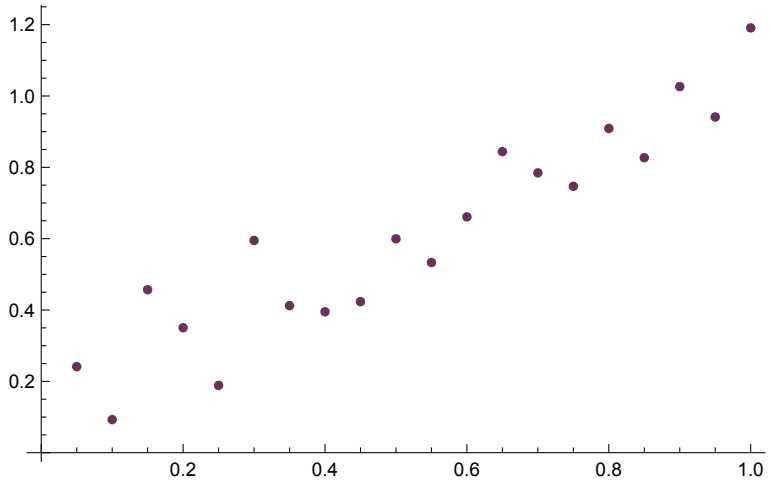
Course organization

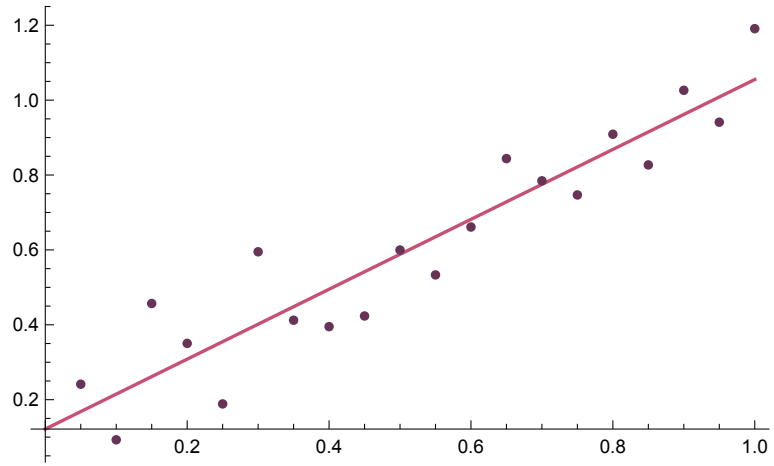
- Course organization
  - Lectures on Zoom: 12 lectures by MPF, 1 lecture by Mattias Villani (SU), and 3 lectures by John Pålsson (If P&C Insurance)
  - Exercise classes on Zoom by Isaac Ren
- Examination
  - 2 projects, groups of 2, deadline same day as exam, pass/fail
  - Exam
- Course literature
  - **Introduction to Linear Regression Analysis** (MPV)
  - An introduction to Statistical Learning (JWHT)
  - Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning (Iz)
  - The Elements of Statistical Learning (HTF)
  - Statistical Learning with Sparsity: The Lasso and Generalizations (HTW)

Todays lecture

- An introduction to simple linear regression
- Estimation of the parameters
- Basic assumptions on the error term
- Properties of the obtained estimators

Motivation





Why use regression?

- Prediction
- Inference

The general regression model

Training data

Assume that we have measurements/training data
$(\mathbf{x}_1, y_1)$, $(\mathbf{x}_2, y_2)$, $\ldots$, $(\mathbf{x}_n, y_n)$.

Regressor variables

$\mathbf{x}_{\cdot 1}$, $\mathbf{x}_{\cdot 2}$, $\ldots$, $\mathbf{x}_{\cdot k}$, are said to be *predictor/regressor variables*.

Response variables

y is said to be a *response variable*.

Goal

We want to describe how the predictor variable(s) can be use to predict the response in the response variable.

Model

$$y = f(\mathbf{x}) + \varepsilon$$

ε is a random variable that captures random errors such as measurement errors or dependencies on variables that are absent in our model.

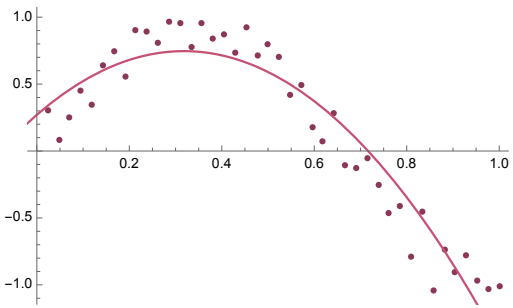
Parametric vs. non-parametric regression

How do we find f?

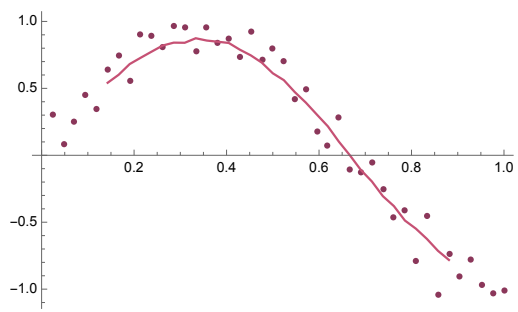
Parametric regression

Assume that f is of a certain form, e.g. a polynomial $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^3$, so that the problem of finding f is reduced to finding the values of the parameters β_0, β_1, β_2, and β_3.

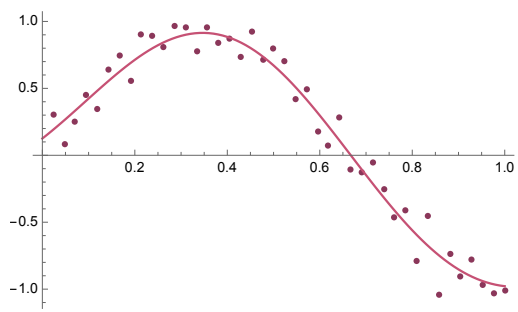
- \+ Possibly to use statistical methods to understand the model
- \+ Will give a function that is easy to describe and use
- – Requires knowledge about the data, or arbitrary guessing of form



Parametric regression
$y = \beta_0 + \beta_1 x + \beta_2 x^2$



Moving average



Kernel regression

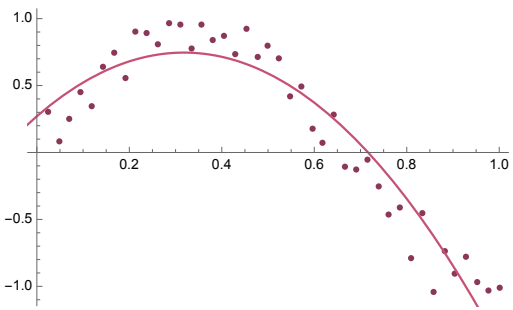
Parametric vs. non-parametric regression

How do we find f?

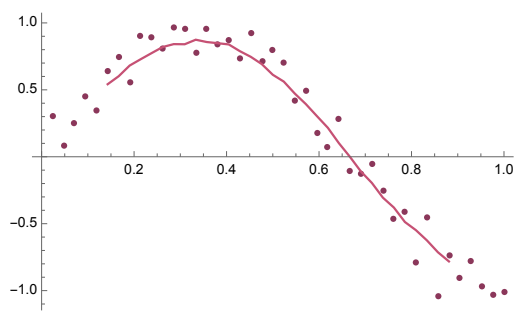
Non-parametric regression

Assume very little about the form of f, and try to find the best choice of f from a very general family, using e.g. splines, moving averages, or kernel regression.

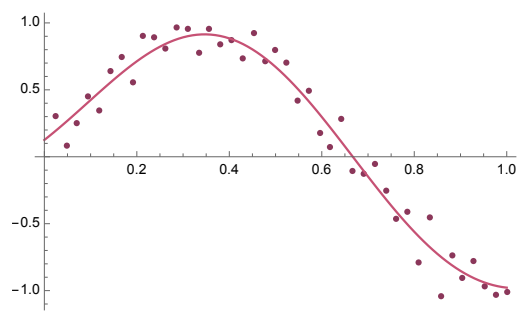
- \+ Can get a very nice fit, even to very complex functions
- \+ Requires only weak assumptions
- – Overfitting
- – Hard to analyze
- – Hard to describe and motivate. Blackbox
- – Usually requires manually setting some kind of "bandwidth"



Parametric regression
$y = \beta_0 + \beta_1 x + \beta_2 x^2$

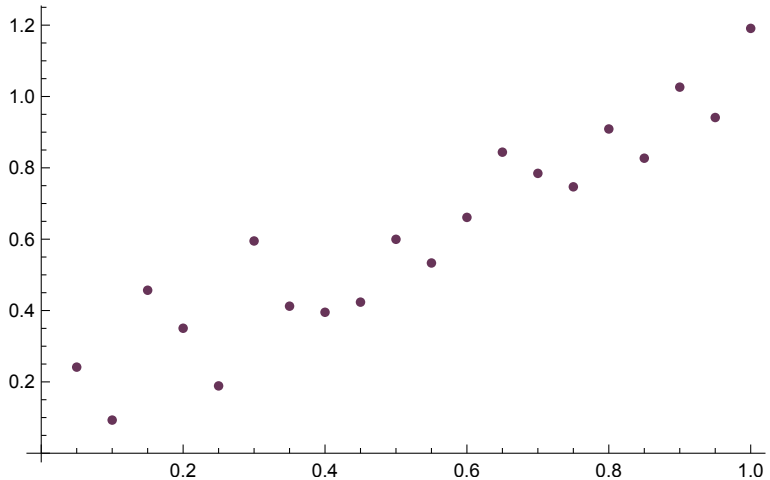


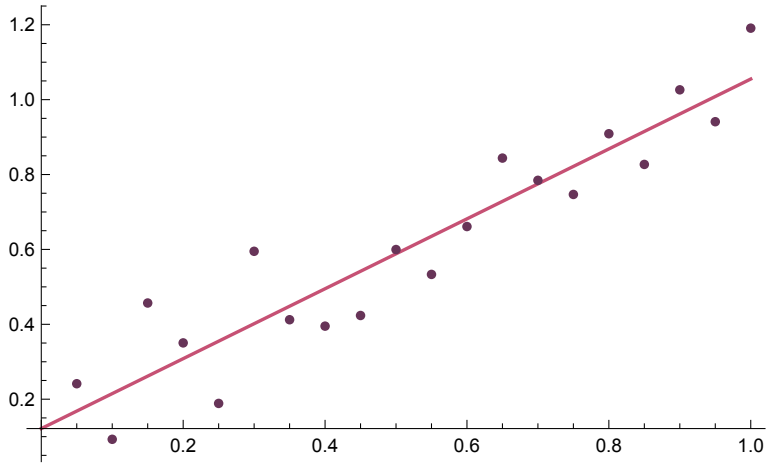
Moving average



Kernel regression

Simple linear regression





Model (simple linear regression)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Regression coefficients

β_0 and β_1 are called *regression coefficients*.

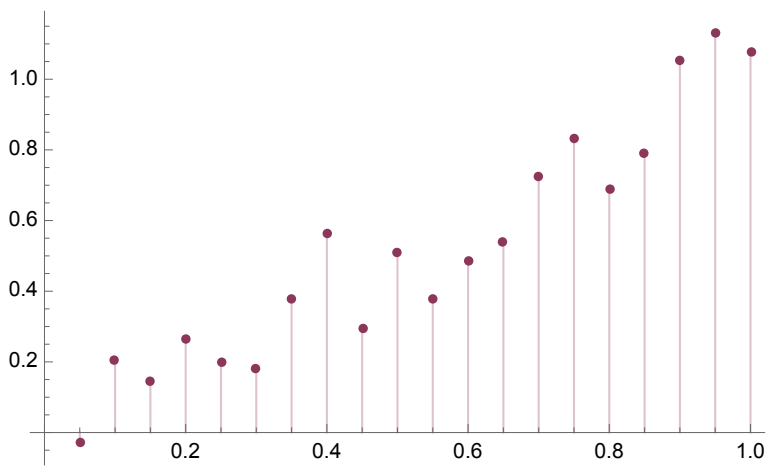
Two types of ways to collect data

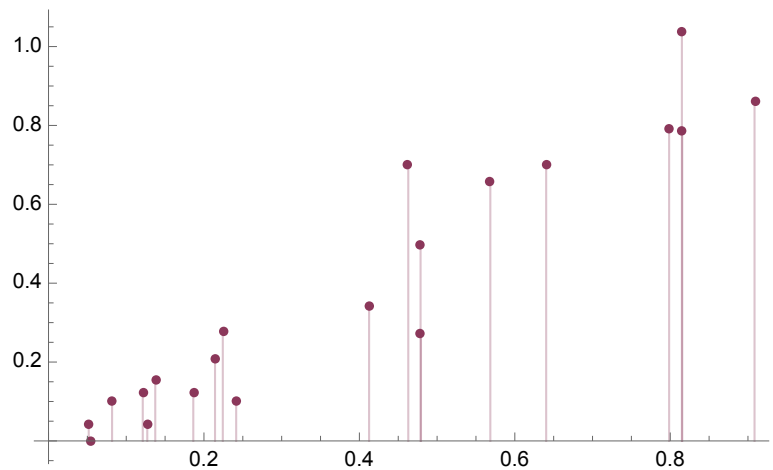
The regression variables $x_1, x_2, \ldots$ can be either

(1) non-random (e.g. data from planned experiments), or

(2) random (common when we use already collected data).

In the lectures, we will initially mostly cover (1), but you will read about (2) in the course literature, and it is often more realistic in practice.



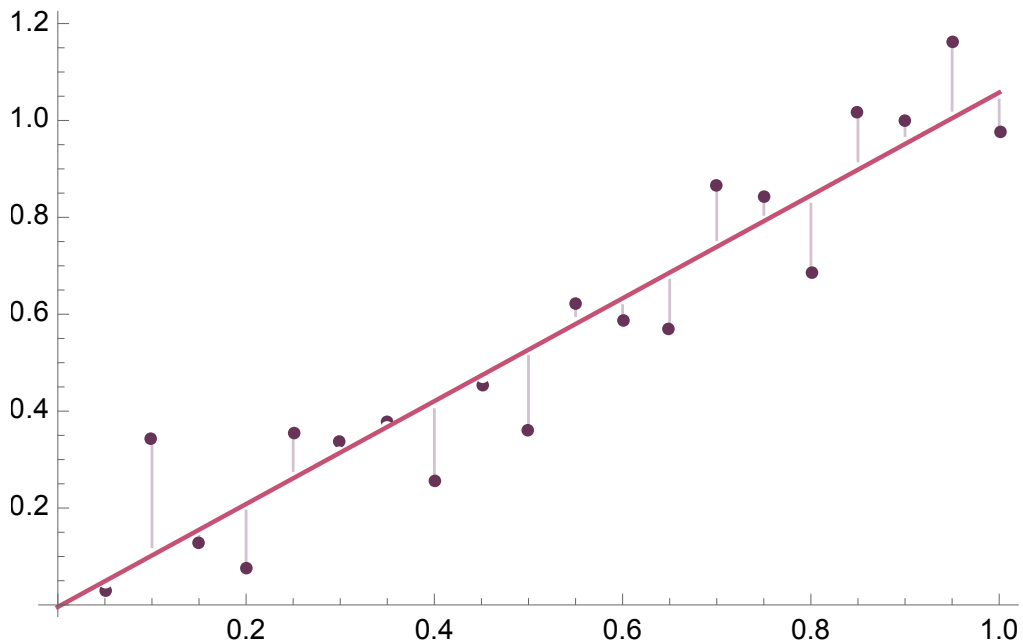


How do we choose β_0 and β_1?

Least squares estimates

Find the pair $(\hat{\beta}_0, \hat{\beta}_1)$ which minimizes the distances between the values $\hat{\beta}_0 + \hat{\beta}_1 x_i$ predicted by the model, and the actual values y_i measured. In other words, we could try and find the coefficients β_0 and β_1 that minimizes

$$f(\beta_0, \beta_1) := \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$



How do we choose β_0 and β_1?

Least squares estimators

$$f(\beta_0, \beta_1) := \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

$$\begin{cases} \frac{d}{d\beta_0} f(\beta_0, \beta_1) = -2 \sum (y_i - (\beta_0 + \beta_1 x_i)) \\ \frac{d}{d\beta_1} f(\beta_0, \beta_1) = -2 \sum x_i (y_i - (\beta_0 + \beta_1 x_i)). \end{cases}$$

If f has a global minimum at $(\beta_0, \beta_1) = (\hat{\beta}_0, \hat{\beta}_1)$, then $\hat{\beta}_0$ and $\hat{\beta}_1$ must satisfy

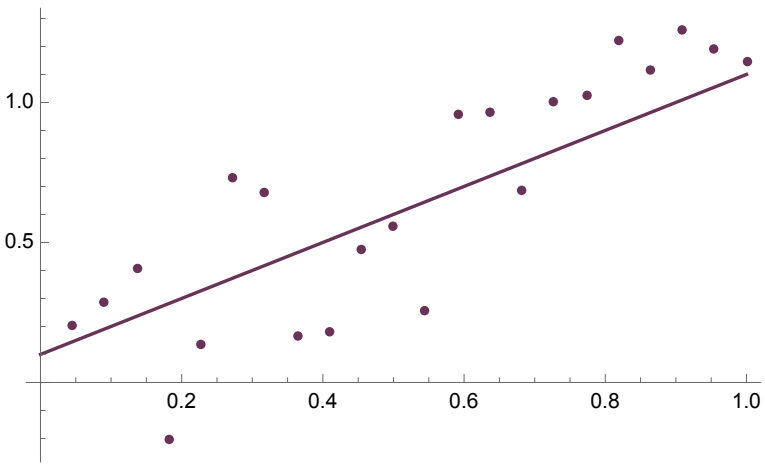
$$\begin{cases} \sum y_i = \hat{\beta}_0 n + \hat{\beta}_1 \sum x_i \\ \sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2. \end{cases}$$

the least squares normal equations

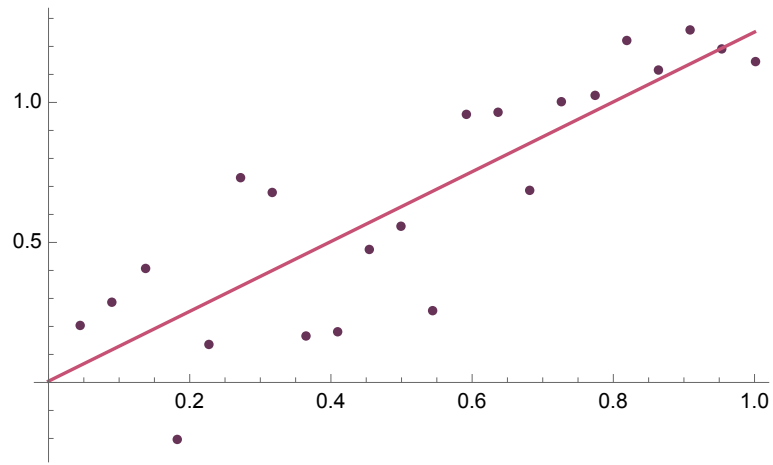
Solving for $\hat{\beta}_0$ and $\hat{\beta}_1$, we see that

$$\begin{cases} \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} \\ \hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \sum x_i \sum x_i} = \frac{n \sum y_i (x_i - \bar{x})}{n \sum (x_i - \bar{x})^2} = \frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} =: \frac{S_{xy}}{S_{xx}}. \end{cases}$$

How do we choose β_0 and β_1?

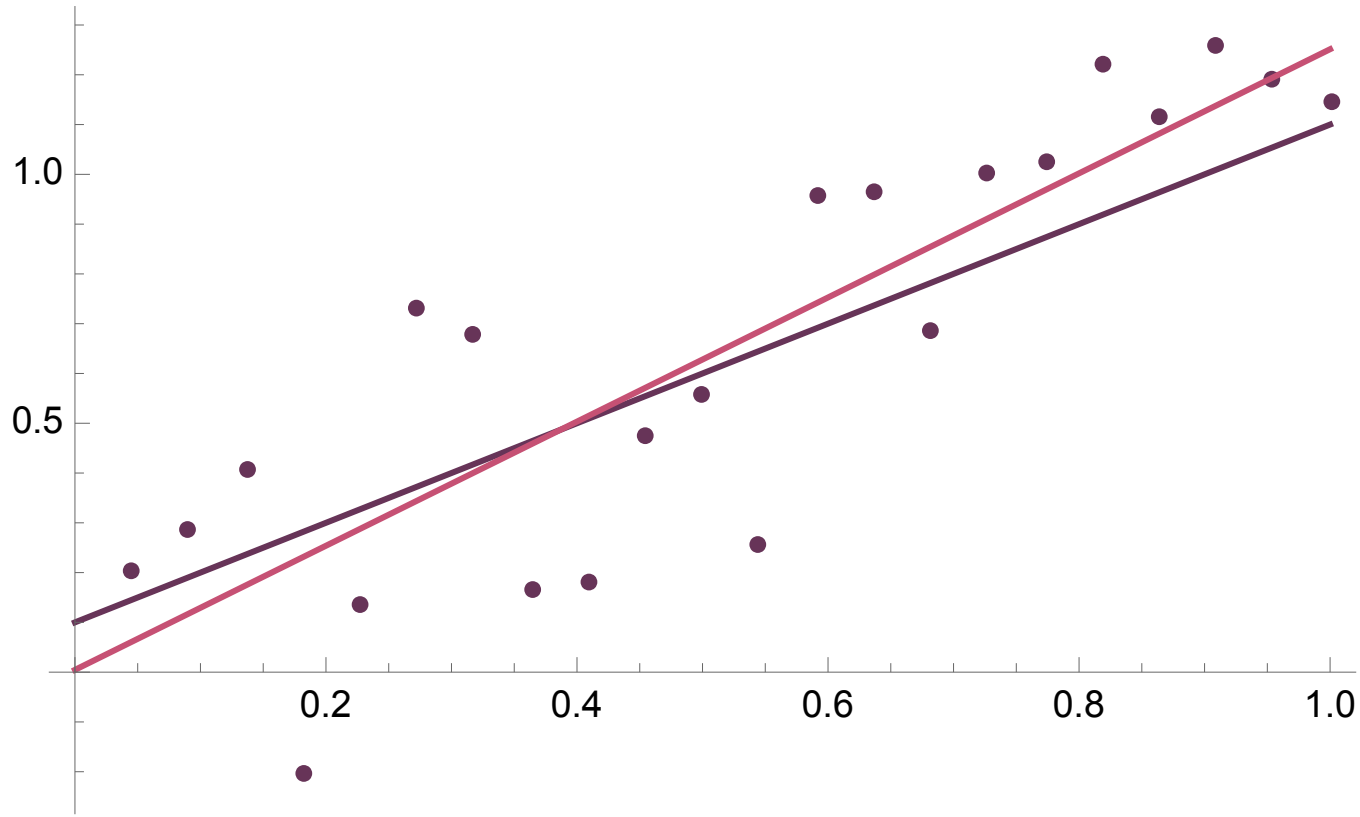


$$y = \beta_0 + \beta_1 x$$



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

How do we choose β_0 and β_1?



$y = \beta_0 + \beta_1 x$ and $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Are the least squares estimates for β_0 and β_1 "good"?

What is a good estimator?

When we estimate parameters from data, we generally want to have the following two properties.

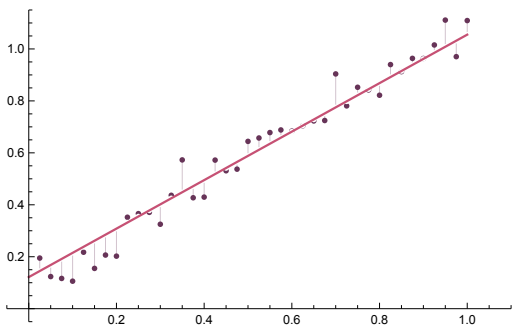
- They should be correct on average (unbiased)
- They should be as close to the true value as possible (small variance)

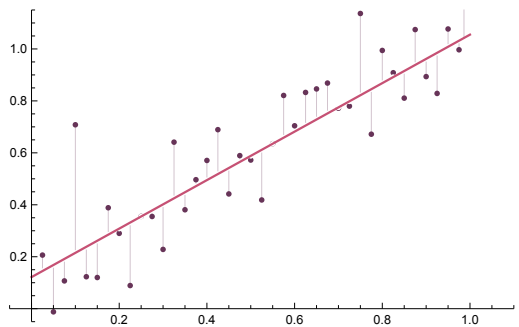
Are the least squares estimates for β_0 and β_1 "good"?

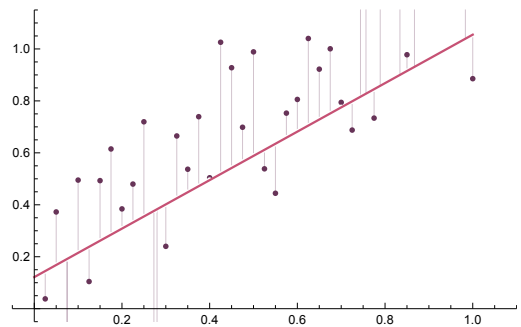
Standard assumptions

To be able to say something about the estimators, we need to make assumptions on ε_i:

1. $\mathbb{E}[\varepsilon_i] = 0$
2. $\mathrm{Var}[\varepsilon_i] = \sigma^2$
3. ε_i and ε_j are independent if $i \neq j$







Note that these assumptions imply that

$$\mathbb{E}[\beta_0 + \beta_1 x_i + \varepsilon_i] = \beta_0 + \beta_1 x_i \qquad \text{and} \qquad \mathrm{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \sigma^2.$$

Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

The expected value of $\hat{\beta}_1$

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\Big[\frac{\sum y_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}\Big] = \frac{\sum \mathbb{E}[y_i](x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{\sum (\beta_0 + \beta_1 x_i)(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$
$$= \frac{\beta_0 \sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} + \frac{\beta_1 \sum x_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = 0 + \beta_1 = \beta_1$$

The expected value of $\hat{\beta}_0$

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}\Big[\frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n}\Big] = \frac{\sum \mathbb{E}[y_i] - \mathbb{E}[\hat{\beta}_1] \sum x_i}{n}$$
$$= \frac{\sum (\beta_0 + \beta_1 x_i) - \beta_1 \sum x_i}{n} = \frac{n\beta_0}{n} = \beta_0$$

Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

The variance of $\hat{\beta}_1$

$$\begin{aligned}\operatorname{Var}(\hat{\beta}_1) &= \operatorname{Var}\left(\frac{\sum y_i(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}\right) = \operatorname{Var}\left(\frac{\sum(\beta_0+\beta_1 x_i+\varepsilon_i)(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}\right) \\ &= \operatorname{Var}\left(\frac{\sum \varepsilon_i(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}\right) = \sigma^2\frac{\sum(x_i-\bar{x})^2}{(\sum(x_i-\bar{x})^2)^2} = \frac{\sigma^2}{\sum(x_i-\bar{x})^2} = \frac{\sigma^2}{S_{xx}}\end{aligned}$$

The variance of $\hat{\beta}_0$

$$\begin{aligned}\operatorname{Var}\hat{\beta}_0 &= \operatorname{Var}(\bar{y}-\hat{\beta}_1\bar{x}) = \operatorname{Var}\bar{y} + \bar{x}^2\operatorname{Var}\hat{\beta}_1 - 2\operatorname{Cov}(\bar{y},\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2\bar{x}^2}{S_{xx}} - 0 = \sigma^2(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}})\end{aligned}$$

Gauss-Markov theorem

Theorem

For simple linear regression, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, and have smallest variance among all other unbiased estimators that are linear in $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$.

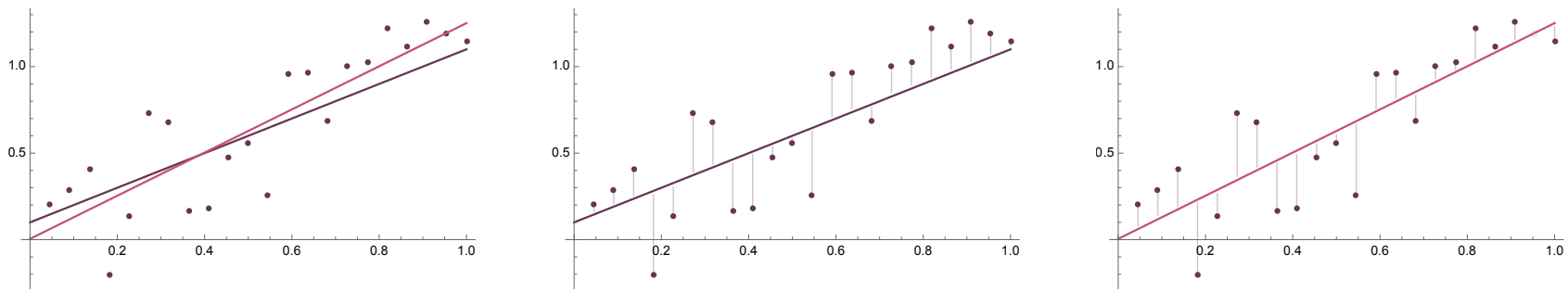
In other words, the least square estimators are the *best linear unbiased estimators*.

How can we estimate σ^2?

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Residuals

If the model is correct, then $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$ should be a i.i.d. sample from a distribution with mean zero and variance σ^2.



Since β_0 and β_1 are unknown, we estimate them with $\hat{\beta}_0$ and $\hat{\beta}_1$ and estimate ε_i by

$$e_i := y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\hat{y}_i}.$$

the ith *residual*

How can we estimate σ^2?

Residual sum of squares

$$
\begin{aligned}
SS_{\text{Res}} &:= \sum e_i^2 = \sum \big(y_i - \hat{y}_i\big)^2 = \sum \big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\big)^2 \\
&= \sum y_i^2 - 2\hat{\beta}_0 n\bar{y} - 2\hat{\beta}_1 \sum y_i x_i + n\hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 n\bar{x} + \hat{\beta}_1^2 \sum x_i^2 \\
&= \sum y_i^2 - 2(\bar{y} - \hat{\beta}_1\bar{x})n\bar{y} - 2\hat{\beta}_1 \sum y_i x_i + n(\bar{y} - \hat{\beta}_1\bar{x})^2 \\
&\quad + 2(\bar{y} - \hat{\beta}_1\bar{x})\hat{\beta}_1 n\bar{x} + \hat{\beta}_1^2 \sum x_i^2 \\
&= \sum y_i^2 - n\bar{y}^2 - 2\hat{\beta}_1 \underbrace{\sum y_i(x_i - \bar{x})}_{S_{xy}} + \hat{\beta}_1 \underbrace{\hat{\beta}_1 \sum x_i(x_i - \bar{x})}_{\frac{S_{xy}}{S_{xx}} \cdot S_{xx} = S_{xy}} \\
&= \underbrace{\sum (y_i - \bar{y})^2}_{SS_T} - \hat{\beta}_1 S_{xy} = SS_T - \hat{\beta}_1 S_{xy}
\end{aligned}
$$

How can we estimate σ^2?

Residual sum of squares

$$SS_{\text{Res}} = SS_T - \hat{\beta}_1 S_{xy}$$

Expected value of SS_{Res}

One can verify that $\mathbb{E}[SS_{\text{Res}}] = (n-2)\sigma^2$.

Residual mean squared

An unbiased estimator for σ^2 is given by

$$\hat{\sigma}^2 := MS_{\text{Res}} := \frac{SS_{\text{Res}}}{n-2}.$$

One can show that MS_{Res} and $\hat{\beta}_1$ are independent.

Standard error of regression

$\sqrt{\hat{\sigma}^2}$ is called the *standard error of regression.*

Properties of the residuals

- $\sum e_i = 0$

$$\underbrace{n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i}_{\text{first normal equation}} \Leftrightarrow \sum \hat{y}_i = \sum y_i \Leftrightarrow \sum e_i = 0$$

- $\sum x_i e_i = 0$

$$\underbrace{\hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i}_{\text{the second normal equation}} \Leftrightarrow \sum x_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \sum x_i y_i$$

$$\Leftrightarrow \sum x_i(\hat{y}_i - y_i) = 0 \Leftrightarrow \sum x_i e_i = 0$$

- $\sum \hat{y}_i e_i = 0$

$$\sum \hat{y}_i e_i = \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i = \hat{\beta}_0 \underbrace{\sum e_i}_{=0} + \hat{\beta}_1 \underbrace{\sum x_i e_i}_{=0} = 0 + 0 = 0$$

The covid dataset

```r
df <- read.csv("/Users/malin/Dropbox/Jobb/Teaching/KTH - SF2930/data.csv", header = TRUE)
```

This dataframe has 135581 rows and 67 columns. Each row contains of one set of data for one country on a specific data, describing things as # new cases that day, total cases to far, etc.
We remove all but the last entry for each country, and then remove the columns that now make little sense.

```r
library(dplyr)

df0 <- df %>% group_by(location) %>% slice(n()) %>% ungroup

df0 <- df0[which(df0$continent!=""),]

df0 <- df0[,c("continent", "location","total_cases_per_million", "total_deaths_per_million","median_age", "gdp_per_capita","hospital_beds_per_thousand", "people_fully_vaccinated_per_hundred","population","aged_65_older","diabetes_prevalence","cardiovasc_death_rate","population_density", "male_smokers","life_expectancy" )]
```

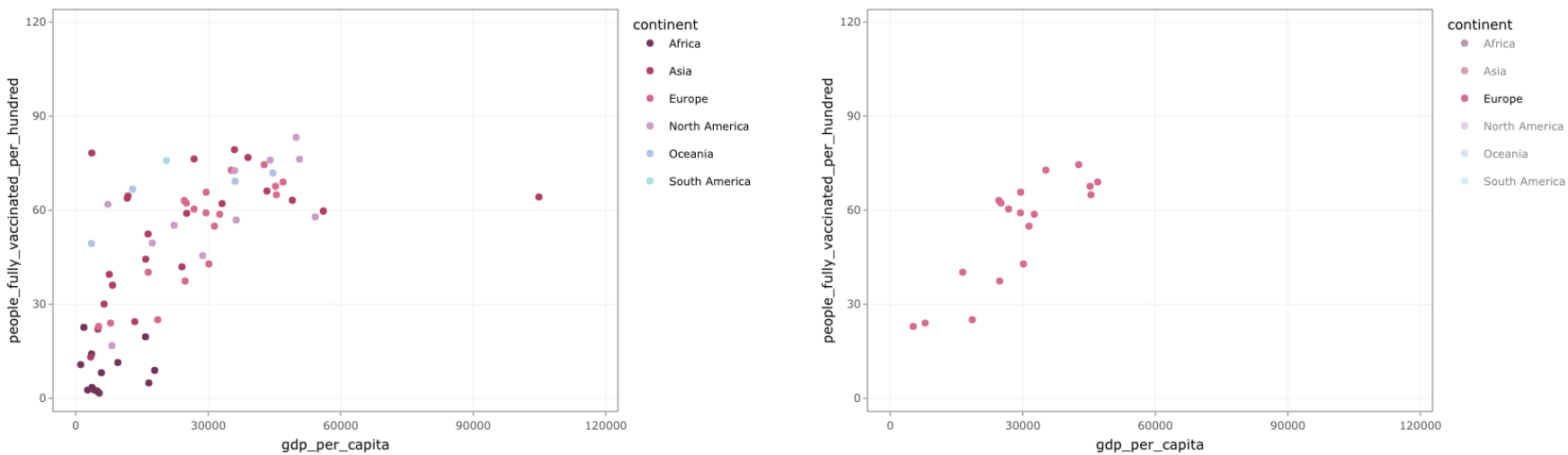
The covid dataset

We now plot the relevant data. ggplotly is useful since it allows us to easily check which country corresponds to each datapoint.

```
library(ggplot2)
library(plotly)

pp <- ggplot(df0, aes(gdp_per_capita, people_fully_
    vaccinated_per_hundred, colour=continent, text=location)
    ) + geom_point()+ scale_color_manual(values=c("white","
    #703457","#AC3F63","#D46B8D", "#C79EC6", "#AFC4E0", "#
    A8DEE1")) + theme_bw()

ggplotly(pp, tooltip="text")
```



The covid dataset

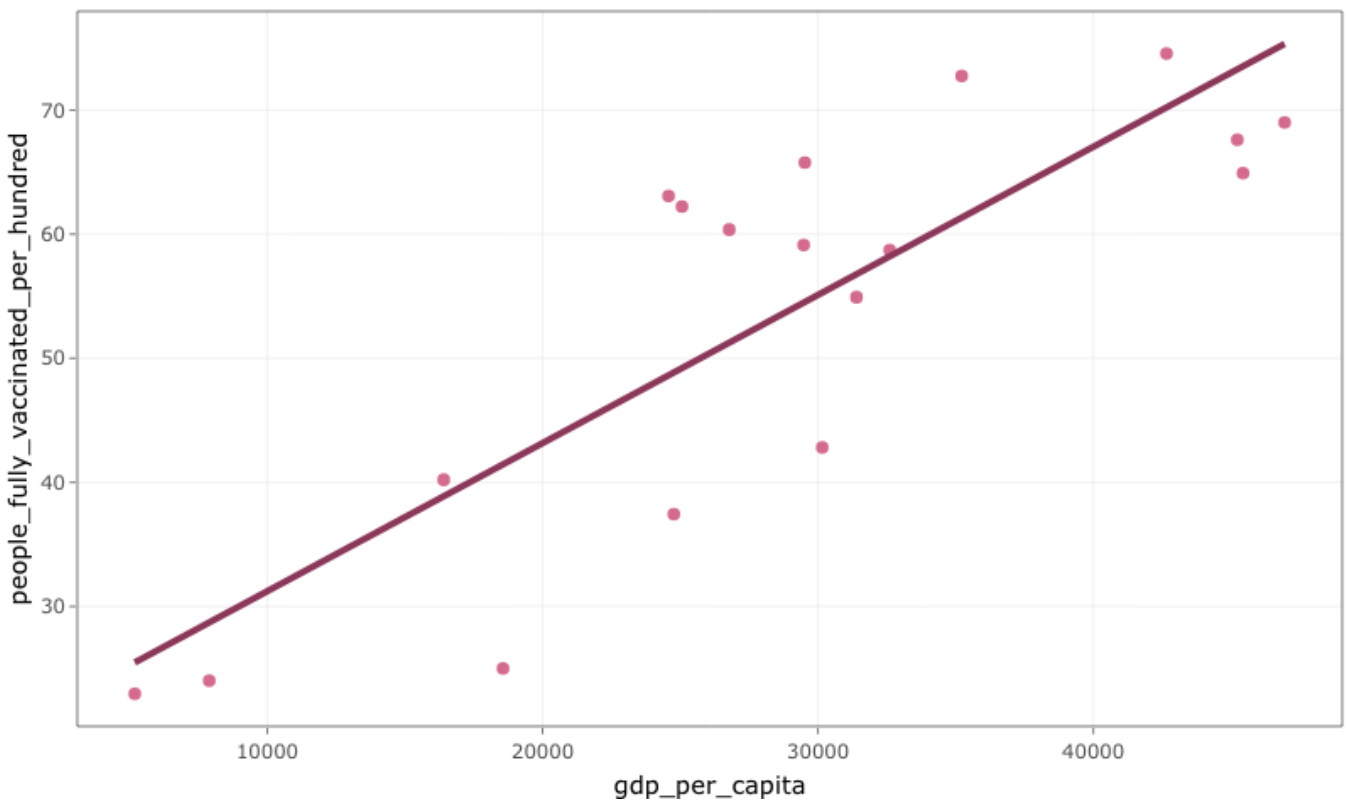
Today we will investigate a potential relationship between the variables `gdp_per_capita` and `people_fully_vaccinated_per_hundred` for countries in Europe where this data is available.

```
df00 <- df0[df0$continent=="Europe",]
df00 <- df00[!is.na(df00[,"gdp_per_capita"]),]
df00 <- df00[!is.na(df00[,"people_fully_vaccinated_per_
    hundred"]),]
```

Example

We now fit a linear model to our data, using gdp_per_capita as a regression variable and people_fully_vaccinated_per_hundred as the response variable.

```
pp <- ggplot(df00, aes(x=gdp_per_capita, y=people_fully_
    vaccinated_per_hundred)) + geom_point(aes(text=location)
    ) + geom_smooth(method=lm, se=FALSE)
ggplotly(pp, tooltip="text")
```



Example

```r
df00.model <- lm(people_fully_vaccinated_per_hundred~gdp_per_capita, data = df00)
summary(df00.model)
```

```
Call:
lm(formula = people_fully_vaccinated_per_hundred ~ gdp_per_
    capita,
    data = df00)

Residuals:
    Min      1Q  Median      3Q     Max
-16.428  -6.176  -0.675   7.997  14.445

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.929e+01  6.075e+00   3.175  0.00588 **
gdp_per_capita 1.194e-03  1.957e-04   6.100 1.53e-05 ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.665 on 16 degrees of freedom
Multiple R-squared:  0.6993,  Adjusted R-squared:  0.6805
F-statistic: 37.21 on 1 and 16 DF,  p-value: 1.534e-05
```

Extra: Useful formulas

$$\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n\bar{x} - n\bar{x} = 0.$$

$$\sum (x_i - \bar{x})^2 = \sum x_i (x_i - \bar{x}) - \bar{x} \underbrace{\sum (x_i - \bar{x})}_{=0} = \sum x_i (x_i - \bar{x}) = S_{xx}$$