Patrice Koehl

Supervised Learning: Linear Regression

Predicting

We get back to the scenario where we would like to predict the value of one variable using another (or a set of other) variables.

Examples:

- Predicting the value of stocks based on its current market
- Predicting the weather next fall based on previous years

Previously, we have seen that we can use k-Nearest Neighbor to do such predictions…
but not always! k-NN will work best when the variable we want to predict
is within the range of the training set…It means that kNN most likely will not work well for prediction outside that range, such as the prediction mentioned above.

Prediction vs Estimation

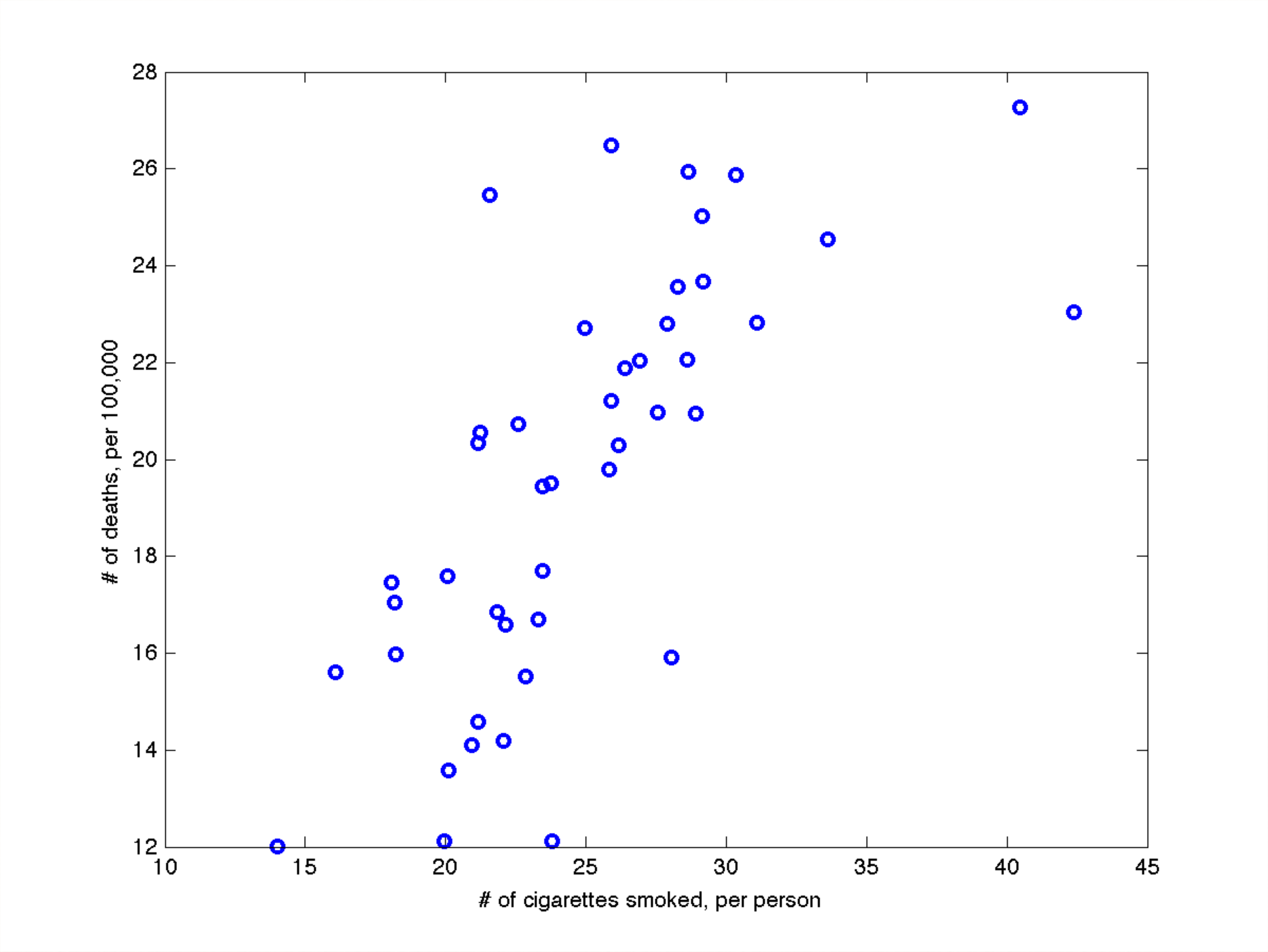
When we use a set of measurements, $(x_{i,1}, \ldots, x_{i,p})$ to predict a value for the response variable, we denote the ***predicted*** value by:

$$\hat{y}_i = \hat{f}(x_{i,1}, \ldots, x_{i,p}).$$

For some problems, we do not care about the specific expression of $\hat{f}$, we just want to make our predictions $\hat{y}$'s as close to the observed values y's as possible. These are called ***prediction problems***. *Example: kNN*

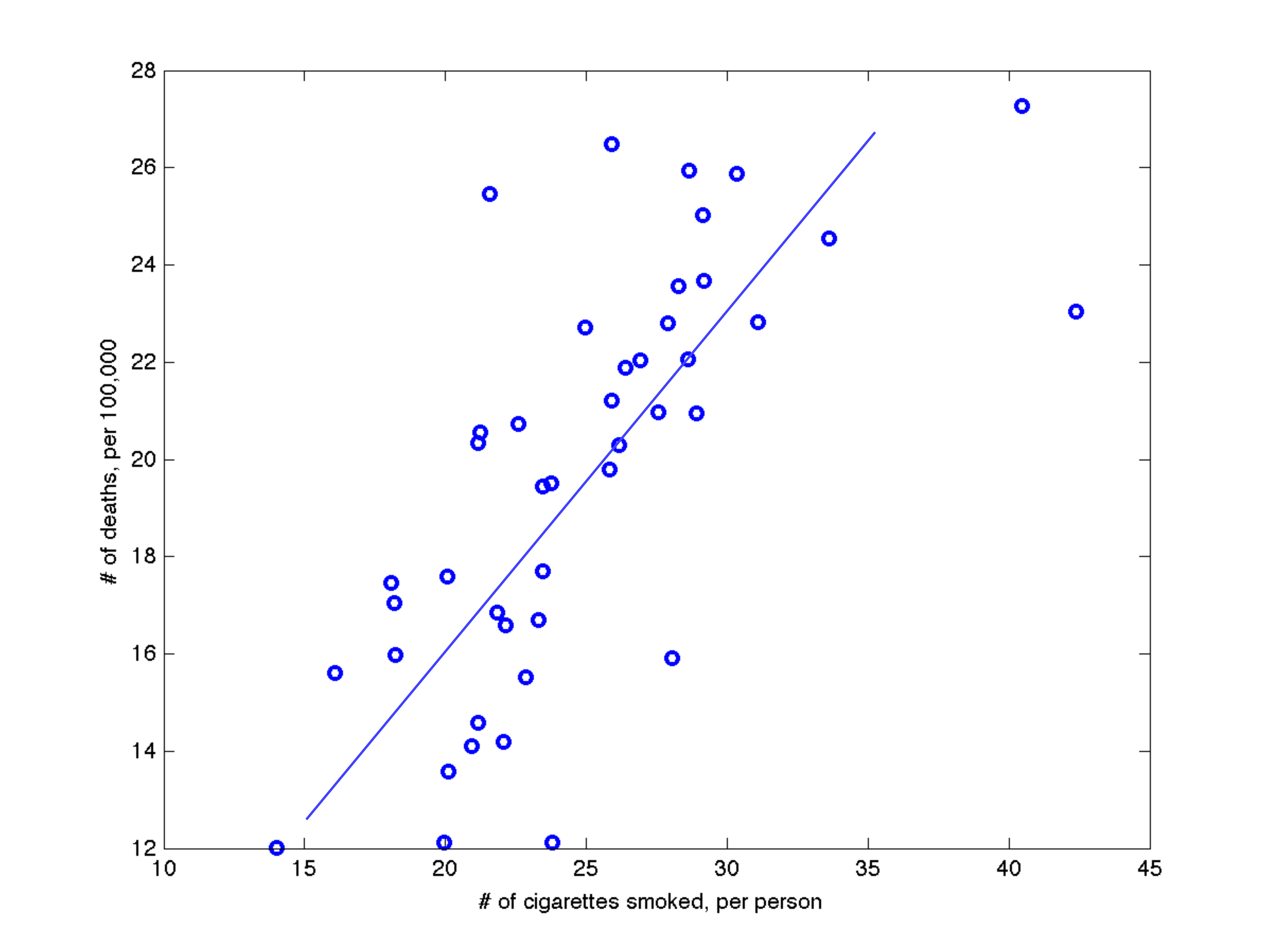
For some problems, what's important is obtaining $\hat{f}$, the estimate of f. These are called ***inference*** problems. *Example: Linear regression*

Linear Regression



Example: let us consider the dataset that gives the number of death observed in a population of smokers

Linear Regression

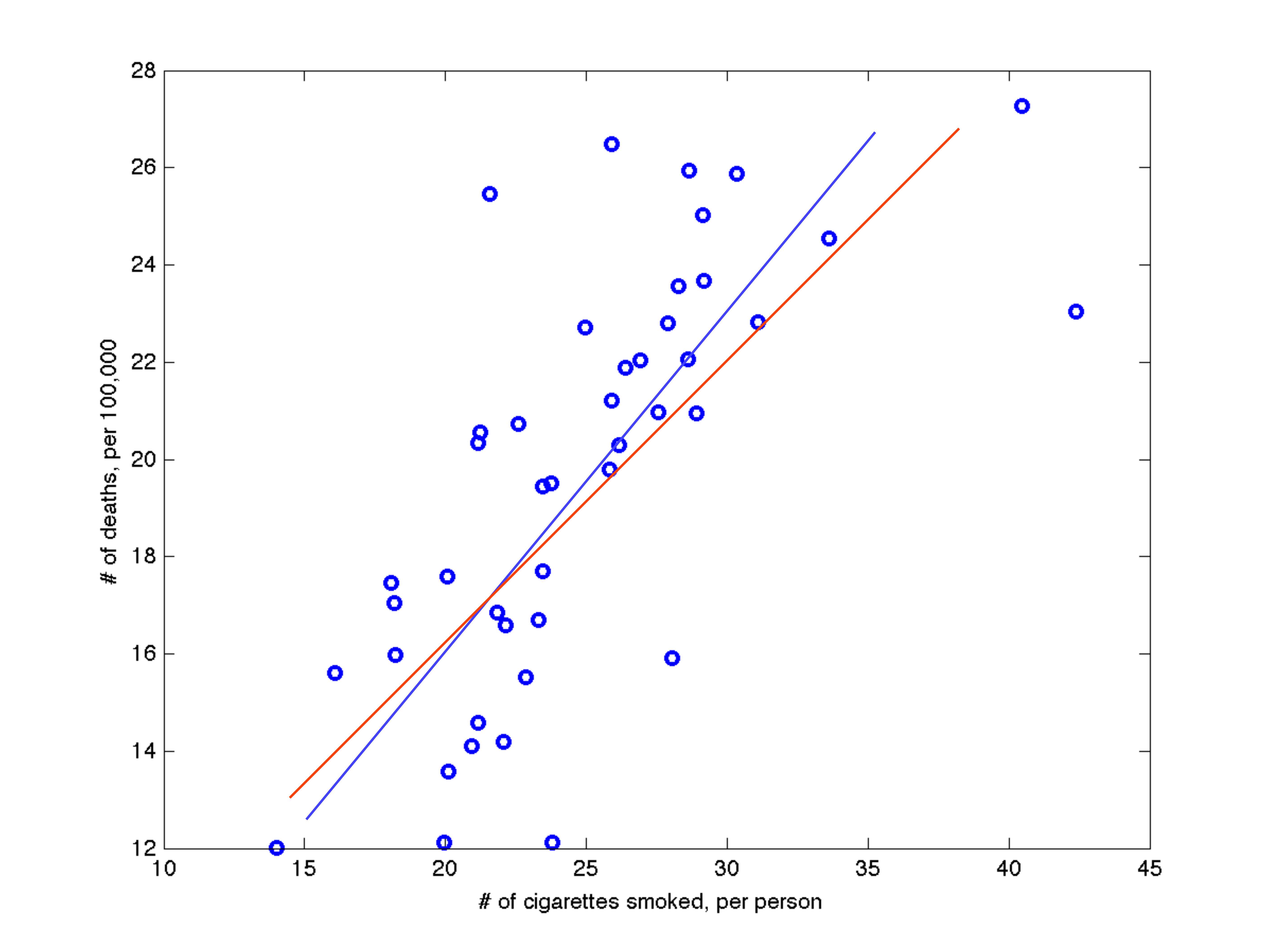


Example: let us consider the dataset that gives the number of death observed in a population of smokers

We assume that these data can be represented with a linear model

Which line is good:
Maybe this one (blue)

Linear Regression

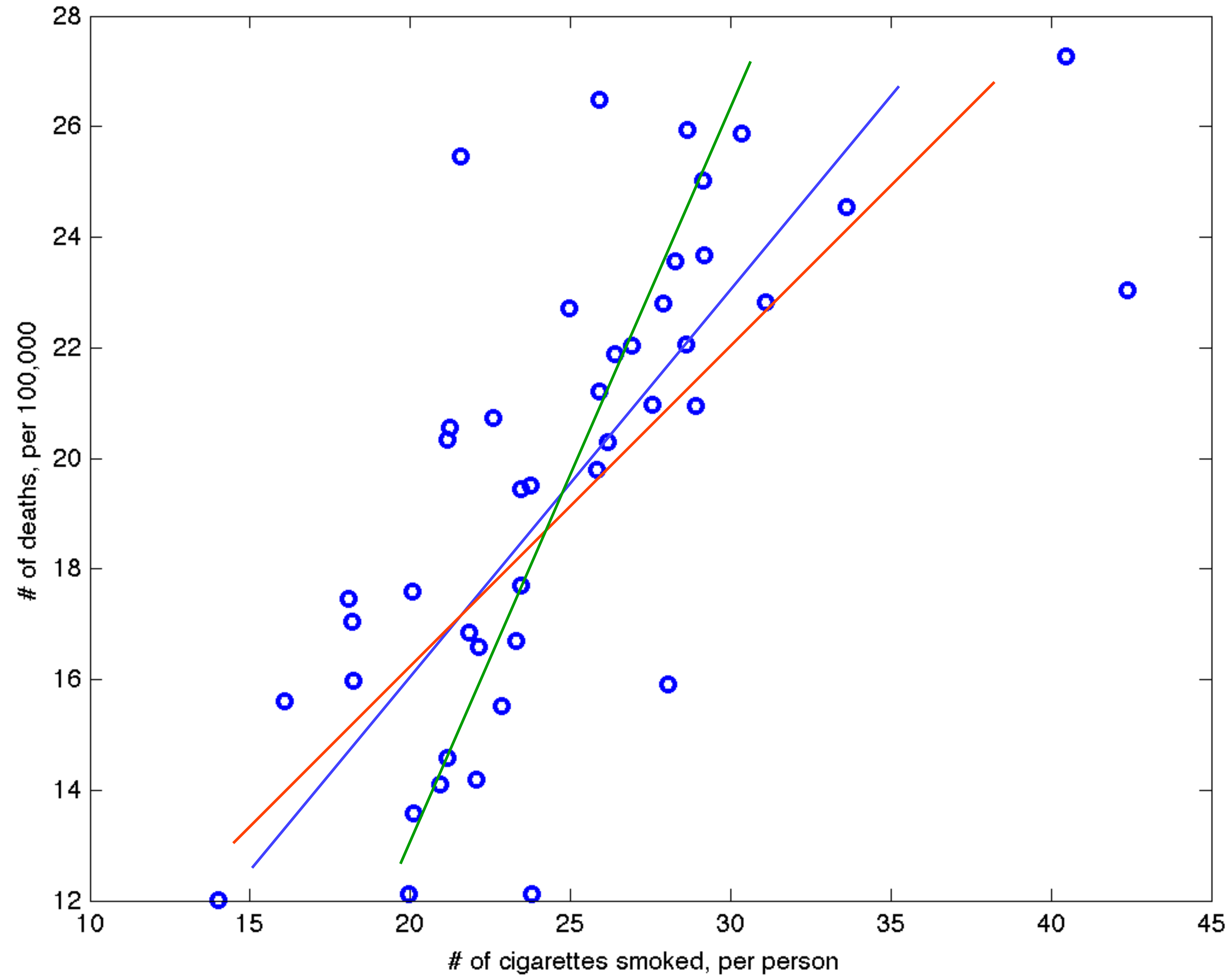


Example: let us consider the dataset that gives the number of death observed in a population of smokers

We assume that these data can be represented with a linear model

Which line is good:
Maybe this one (blue)
Or this one (red)

Linear Regression



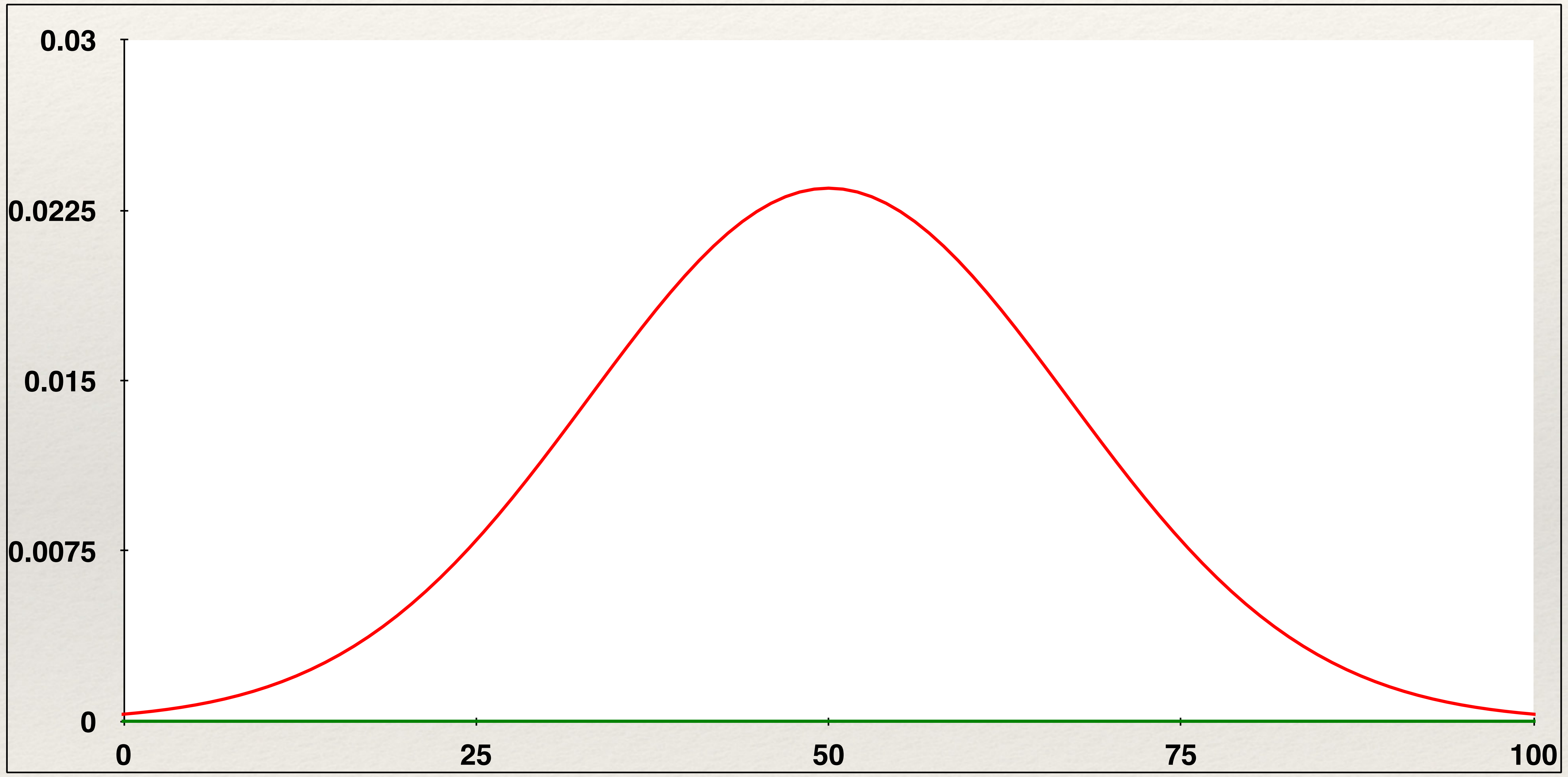
Example: let us consider the dataset that gives the number of death observed in a population of smokers

We assume that these data can be represented with a linear model

Which line is good:
Maybe this one (blue)
Or this one (red)
Or this one (green)

The normal distribution

In everyday life many variables such as height, weight, shoe size and exam marks all tend to be normally distributed, that is, they all tend to look like:

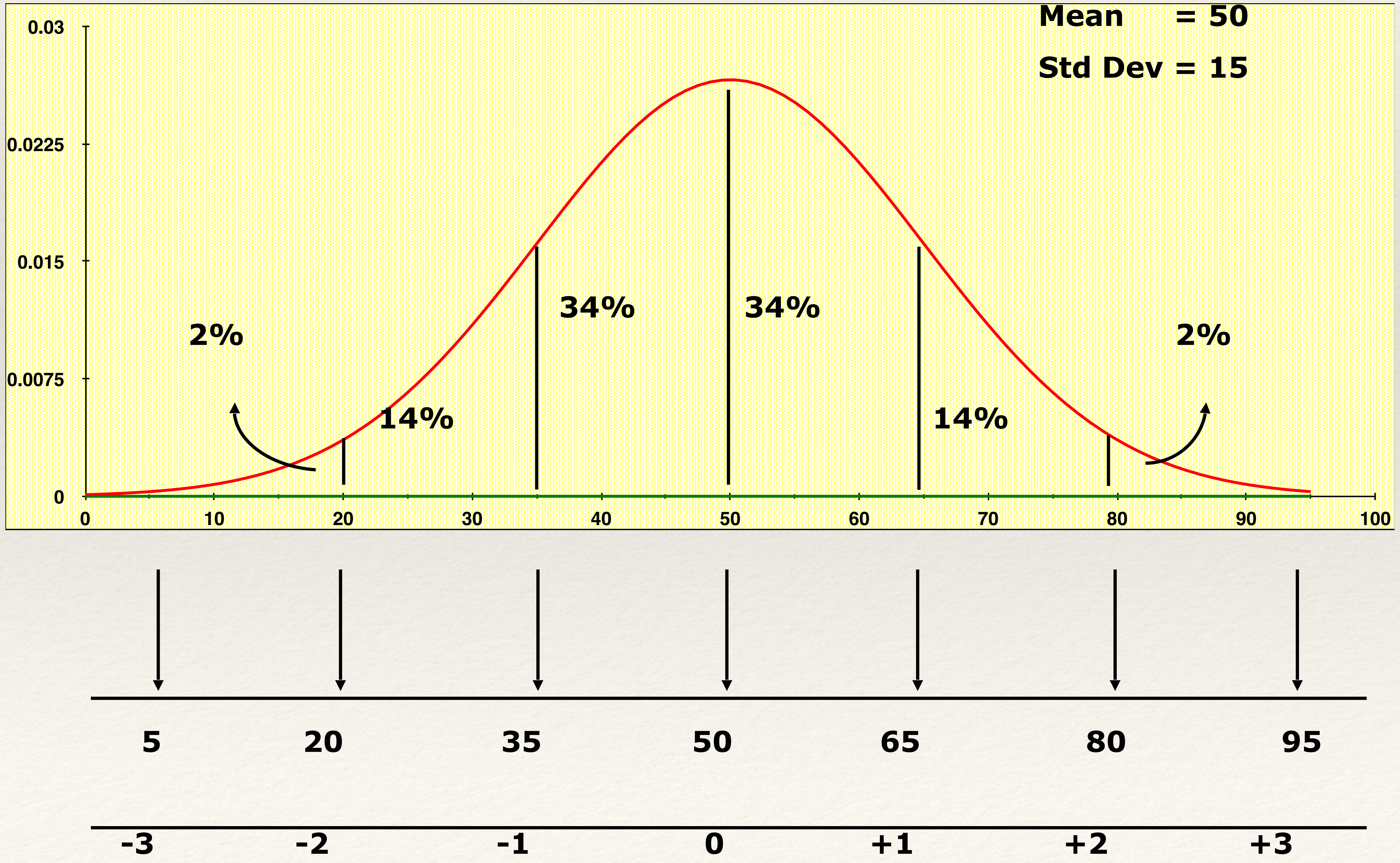


It is bell-shaped and symmetrical about the mean

The mean, median and mode are equal

The normal distribution

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}}$$



Linear Regression

Let us suppose that:

- The data points are independent of each other
- Each data point has a measurement error that is random, distributed as a Normal distribution around the "true" value $Y(x_i)$:

$$f(y_i;Y) = \exp\left[-\frac{1}{2}\left(\frac{y_i - Y(x_i)}{\sigma_i}\right)^2\right]$$

The likelihood function is:

$$L(Y) = f(y_1,\ldots,y_N;Y) \approx f(y_1;Y)\ldots f(y_N;Y)$$

$$L(Y) = \prod_{i=1}^{N}\left\{\exp\left[-\frac{1}{2}\left(\frac{y_i - Y(x_i)}{\sigma_i}\right)^2\right]\right\}$$

Linear Regression

Let us suppose that:

➢The data points are independent of each other

➢Each data point has a measurement error that is random, distributed as a Normal distribution around the "true" value $Y(x_i)$:

The probability of the data points, *given the model Y* is then:

$$P(data/Model) \propto \prod_{i=1}^{N} \left\{ \exp\left[-\frac{1}{2} \left(\frac{y_i - Y(x_i)}{\sigma_i} \right)^2 \right] \right\}$$

Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where *A* and *B* are events and $P(B) \neq 0$.

- ***P(A | B)*** is a conditional probability: it is the probability of event ***A*** occurring given that ***B*** is true. It is also called the posterior probability of ***A*** given ***B***.

- ***P(B | A)*** is also a conditional probability: the probability of event ***B*** occurring given that ***A*** is true.

- ***P(A)*** and ***P(B)*** are the probabilities of observing A and B respectively without any given conditions; they are known as the marginal probability or prior probability.

- ***A*** and ***B*** must be different events.

Bayes' theorem

User new evidence to update beliefs

Likelihood function

Prior probability

$$P(Model\ /\ Data) = \frac{P(Data\ /\ Model)\,P(Model)}{P(Data)}$$

Posterior probability

Model evidence (Independent of Model)

Bayes' Theorem

Hypothesis (model) on your friend's new baby:

H1: brown hair baby boy

H2: blond hair baby girl

H3: cute baby cat

Bayes' Theorem

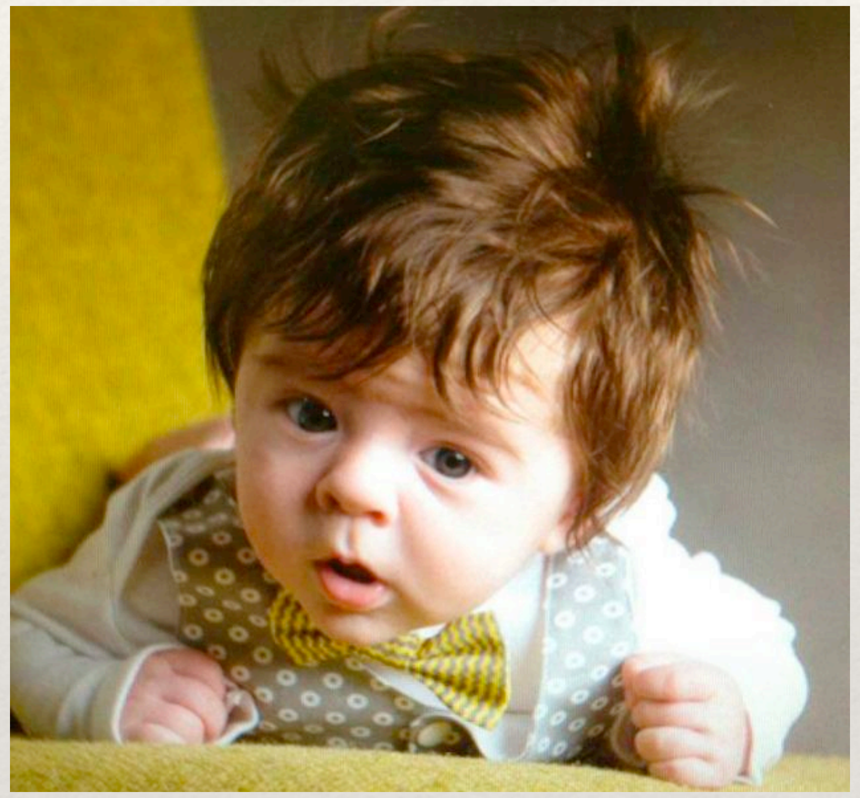
Hypothesis (model) on your friend's new baby:

H1: brown hair baby boy

H2: blond hair baby girl

H3: cute baby cat

Evidence



Bayes' Theorem

Hypothesis (model) on your friend's new baby:

H1: brown hair baby boy

H2: blond hair baby girl

H3: cute baby cat

Evidence

Bayes' Theorem

Example: suppose a drug test is 99% sensitive and 99% specific.
(Namely, $P(+|User) = 0.99$ and $P(+|Non\ user) = 0.01$)

Suppose that 0.5% of people are users of the drug. If a random individual tests positive, what is the probability she is a user?

Bayes' Theorem

Example: suppose a drug test is 99% sensitive and 99% specific.
(Namely, $P(+|User) = 0.99$ and $P(+|Non\ user) = 0.01$)

Suppose that 0.5% of people are users of the drug. If a random individual tests positive, what is the probability she is a user?

$$P(User\,|\,+) = \frac{P(+\,|\,User)P(User)}{P(+)} = \frac{P(+\,|\,User)P(User)}{P(+\,|\,User)P(User) + P(+\,/\,NonUser)P(NonUser)}$$

$$P(User\,|\,+) = 33.2\%$$

Linear Regression

The probability of the data points, *given the model Y* is then:

$$P(data/Model) \propto \prod_{i=1}^{N} \left\{ \exp\left[-\frac{1}{2} \left(\frac{y_i - Y(x_i)}{\sigma_i} \right)^2 \right] \right\}$$

Linear Regression

The probability of the data points, *given the model Y* is then:

$$P(data/Model) \propto \prod_{i=1}^{N}\left\{\exp\left[-\frac{1}{2}\left(\frac{y_i - Y(x_i)}{\sigma_i}\right)^2\right]\right\}$$

Application of Bayes 's theorem:

$$P(Model/Data) \propto P(Data/Model)P(Model)$$

Linear Regression

The probability of the data points, *given the model Y* is then:

$$P(data/Model) \propto \prod_{i=1}^{N} \left\{ \exp\left[-\frac{1}{2}\left(\frac{y_i - Y(x_i)}{\sigma_i} \right)^2 \right] \right\}$$

Application of Bayes 's theorem:

$$P(Model/Data) \propto P(Data/Model)P(Model)$$

With no information on the models, we can assume that the prior probability P(Model) is constant.

Linear Regression

The probability of the data points, *given the model Y* is then:

$$P(data/Model) \propto \prod_{i=1}^{N} \left\{ \exp\left[-\frac{1}{2} \left(\frac{y_i - Y(x_i)}{\sigma_i} \right)^2 \right] \right\}$$

Application of Bayes ‘s theorem:

$$P(Model/Data) \propto P(Data/Model)P(Model)$$

With no information on the models, we can assume that the prior probability P(Model) is constant.

Finding the model that maximizes P(Model/Data) is then equivalent to finding the model that maximizes P(Data/Model).

Linear Regression

The probability of the data points, *given the model Y* is then:

$$P(data/Model) \propto \prod_{i=1}^{N}\left\{\exp\left[-\frac{1}{2}\left(\frac{y_i - Y(x_i)}{\sigma_i}\right)^2\right]\right\}$$

Application of Bayes 's theorem:

$$P(Model/Data) \propto P(Data/Model)P(Model)$$

With no information on the models, we can assume that the prior probability P(Model) is constant.

Finding the model that maximizes P(Model/Data) is then equivalent to finding the model that maximizes P(Data/Model).

This is equivalent to maximizing its logarithm, or minimizing the negative of its logarithm, namely:

$$\chi_2 = \sum_{i=1}^{N}\frac{1}{2}\left(\frac{y_i - Y(x_i)}{\sigma_i}\right)^2$$

Linear Regression

Fitting to a straight line:

$$Y(x) = ax + b$$

Then:

$$\chi_2 = \sum_{i=1}^{N} \left(\frac{y_i - ax_i - b}{\sigma_i} \right)^2$$

The parameters *a* and *b* are obtained from the two equations:

$$\frac{\delta \chi_2}{\delta a} = -2 \sum_{i=1}^{N} \frac{x_i(y_i - ax_i - b)}{\sigma_i^2} = 0$$

$$\frac{\delta \chi_2}{\delta b} = -2 \sum_{i=1}^{N} \frac{y_i - ax_i - b}{\sigma_i^2} = 0$$

Linear Regression

Let us define:

$$S = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \quad S_x = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \quad S_y = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \quad S_{xx} = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \quad S_{xy} = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}$$

Then:

$$aS_{xx} + bS_x = S_{xy}$$
$$aS_x + bS = S_y$$

From which we find a and b:

$$a = \frac{S_{xy}S - S_x S_y}{S_{xx}S - S_x^2}$$

$$b = \frac{S_{xx}S_y - S_x S_{xy}}{S_{xx}S - S_x^2}$$

Linear Regression

We are not done!

Uncertainty on the values of a and b:

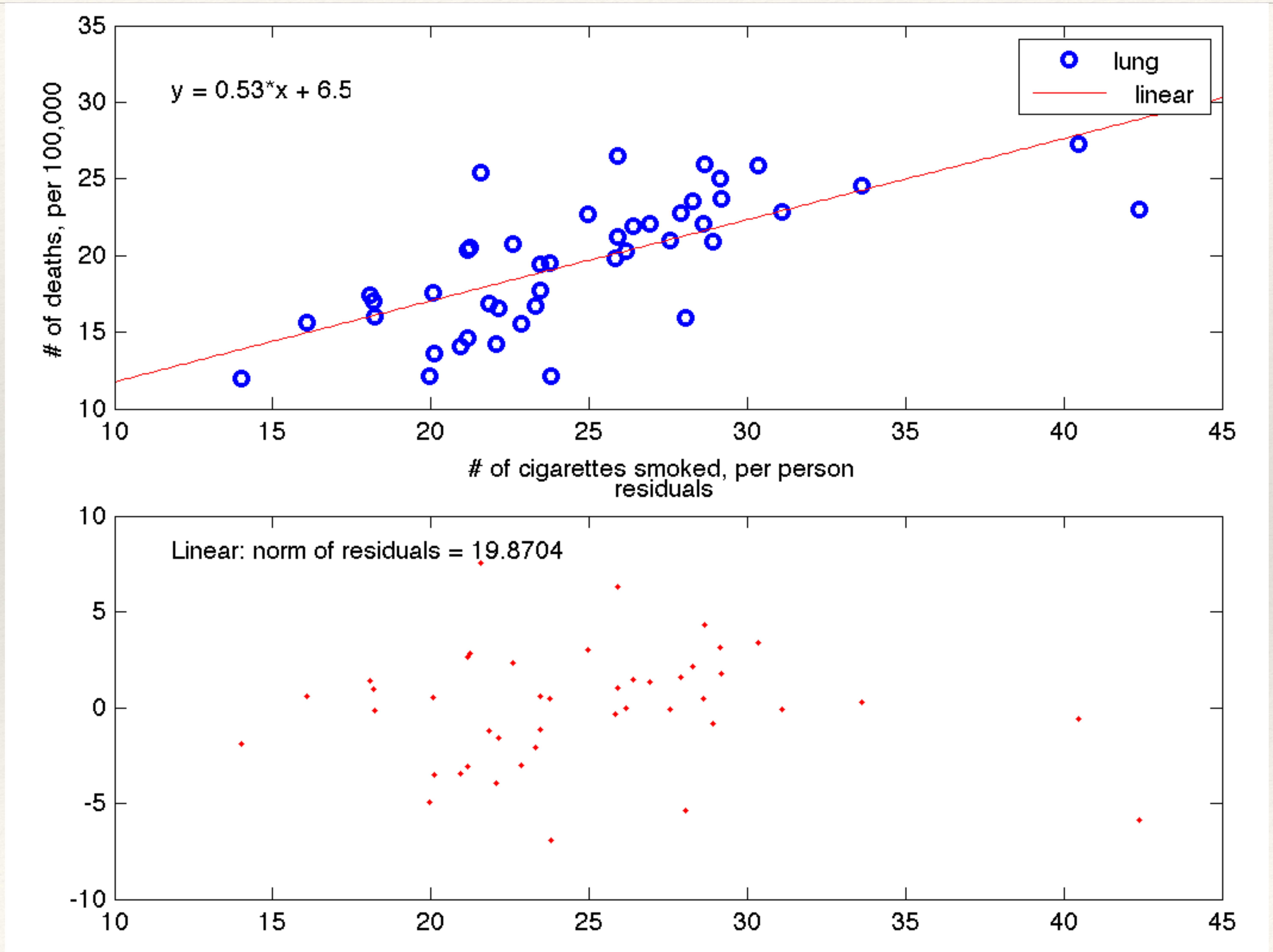
$$\sigma_a^2 = \frac{S}{S_{xx}S - S_x^2}$$

$$\sigma_b^2 = \frac{S_x}{S_{xx}S - S_x^2}$$

Evaluate goodness of fit:

- Compute residual error on each data point: $Y(x_i)\text{-}y_i$
- Compute correlation coefficient R^2

Linear Regression



Linear Regression

More general linear model:

$$Y(x) = a_1 X_1(x) + a_2 X_2(x) + \ldots + a_M X_M(x)$$

Then:

$$\chi_2 = \sum_{i=1}^{N} \left(\frac{y_i - a_1 X_1(x) - a_2 X_2(x) - \ldots - a_M X_M(x)}{\sigma_i} \right)^2$$

The parameters *a* and *b* are obtained from the two equations:

$$\frac{\delta \chi_2}{\delta a_k} = -2 \sum_{i=1}^{N} \frac{X_k(x_i)(y_i - a_1 X_1(x) - a_2 X_2(x) - \ldots - a_M X_M(x))}{\sigma_i^2} = 0$$

Model fitting

Let us work out a simple example. Let us consider we have N students, $S_1, \ldots, S_N$ and let us "evaluate" a variable x_i for each student such that:

$x_i = 1$ if student S_i owns a Ferrari, and $x_i = 0$ otherwise.

We want an estimator of the probability p that a student owns a Ferrari.

The probability of observing x_i for student S_i is given by:

$$f(x_i, p) = p^{x_i}(1 - p)^{1-x_i}$$

The likelihood of observing the values x_i for all N students is:

$$L(p) = f(x_1, x_2, \ldots, x_N; p) \approx f(x_1, p) f(x_2, p) \ldots f(x_N, p)$$

Model Fitting

$$L(p) = p^{\sum x_i}(1-p)^{N-\sum x_i}$$

The maximum likelihood estimator of p is the value p_m that maximizes L(p):

$$p_m = \mathrm{argmax} L(p)$$

This is equivalent to maximizing the logarithm of L(p) (log-likelihood):

$$\log(L(p)) = \log(p)\sum_{i=1}^{N} x_i + \log(1-p)\left(N - \sum_{i=1}^{N} x_i\right)$$

Model Fitting

$$\frac{\delta \log(L(p))}{\delta p} = \frac{1}{p}\sum_{i=1}^{N} x_i - \frac{1}{1-p}\left(N - \sum_{i=1}^{N} x_i\right) = 0$$

$$p_m = \frac{1}{N}\sum_{i=1} x_i$$

This is the most intuitive value…. And it matches with the maximum likelihood estimator