Modelli di regressione

Ilaria Gandin

Corso per le Scuole di Specialità

26 Gennaio 2023

Example I

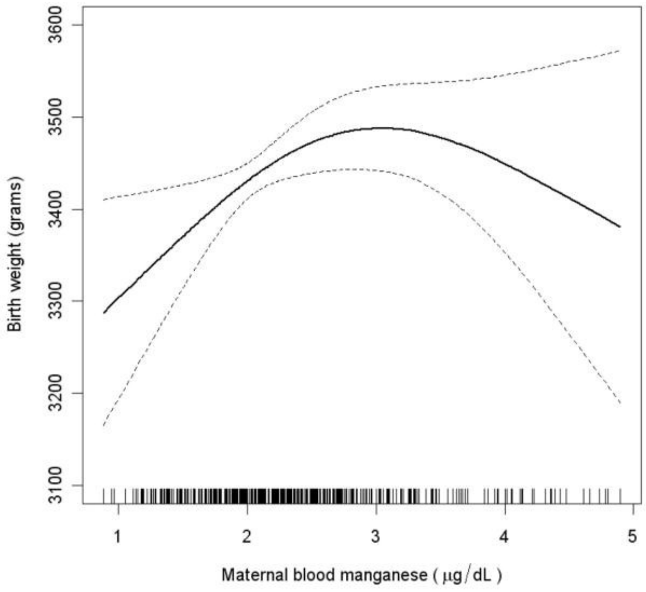
- **Atherosclerotic Cardiovascular Disease** Risk Calculator to determine **10-year risk** of heart disease or stroke
- http://static.heart.org/riskcalc/app/index.html#!/baseline-risk

Example II

Epidemiology. 2009 May ; 20(3): 367–373. doi:10.1097/EDE.0b013e31819b93c0.

Maternal Blood Manganese Levels and Infant Birth Weight

Ami R. Zota[a,b], Adrienne S. Ettinger[a,c,d], Maryse Bouchard[a], Chitra J. Amarasiriwardena[a,c],



“The objective of the present analysis was to examine the relationship between in utero manganese exposure and birth weight”

“Birth weight increased with manganese levels up to 3.1 μg/L, and then a slight reduction in weight was observed at higher levels” app.wooclap.com/OOFUPI

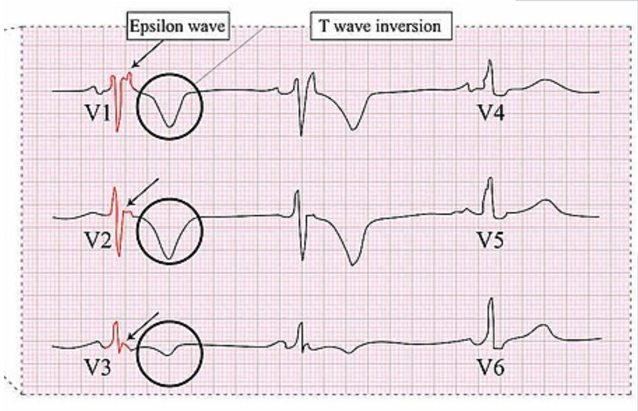
Example III

The American Journal of Cardiology

T-Wave Inversion, QRS Duration, and QRS/T Angle as Electrocardiographic Predictors of the Risk for Sudden Cardiac Death

Jari Antero Laukkanen, MD, PhD[a,b,*], Emanuele Di Angelantonio, MD, PhD[c], Hassan Khan, MD, PhD[c],

“Cox proportional hazards models were used to evaluate the risk of SCD first for TWI [...] with **multivariable adjustment** for **age and clinical factors** (age, alcohol consumption, cigarette smoking, serum low- and high-density lipoprotein cholesterol, systolic blood pressure, type 2 diabetes, BMI, high-sensitivity C-reactive protein, previous myocardial infarction, and cardiorespiratory fitness)”



Outline

- Purpose of regression models
- Simple linear regression
- Multivariable approach
- Logistic regression
- Model building

Purpose of regression models

- **Prediction**: predicting responses of individual subjects
- **Estimation**: estimate the shape and magnitude of the relationship between a predictor variable and a response variable
- **Hypothesis testing**: study association between predictor variable and a response variable after adjusting for the effect of other predictors

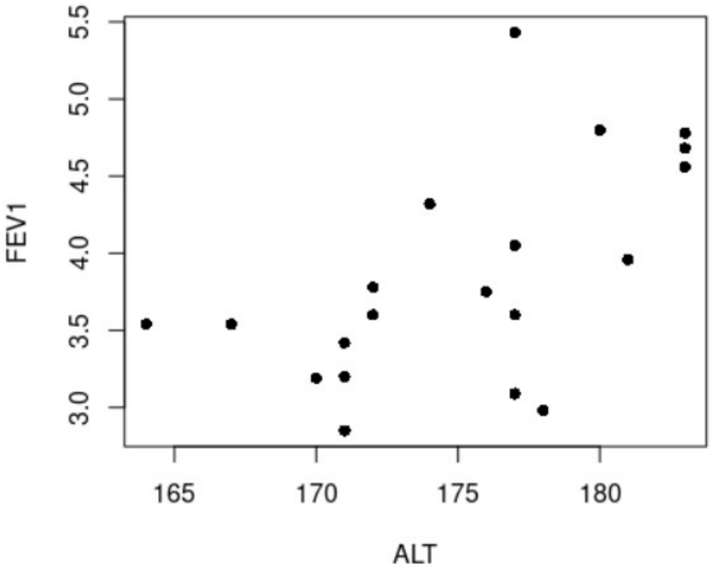
prediction

description

Simple linear regression

<u>Interest</u>: association between height and FEV1

- Response: Y = *FEV1*
- Predictor: X = *height*



ID	Height (cm)	FEV1 (liters)
s1	164.0	3.54
s2	167.0	3.54
s3	170.4	3.19
s4	171.2	2.85
s5	171.2	3.42
s6	171.3	3.20
s7	172.0	3.60
s8	172.0	3.78
s9	174.0	4.32
s10	176.0	3.75
s11	177.0	3.09

Simple linear regression

<u>Interest</u>: association between height and FEV1

- Response: Y = *FEV1*
- Predictor: X = *height*

a and *b* are **coefficients** to be estimated

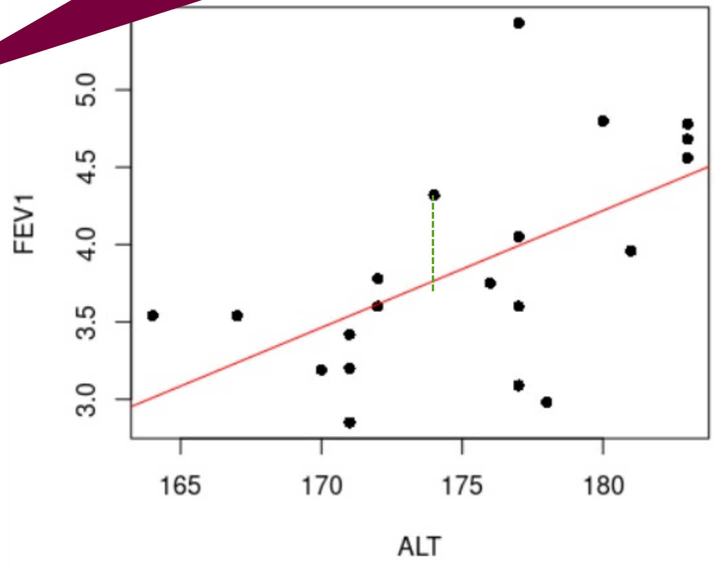
$$Y = a + b \cdot X + E$$

For each subject *i*:

$$y_i = a + b \cdot x_i + e_i$$

For subject s2:

$$3.54 = a + b \cdot 167 + e_2$$

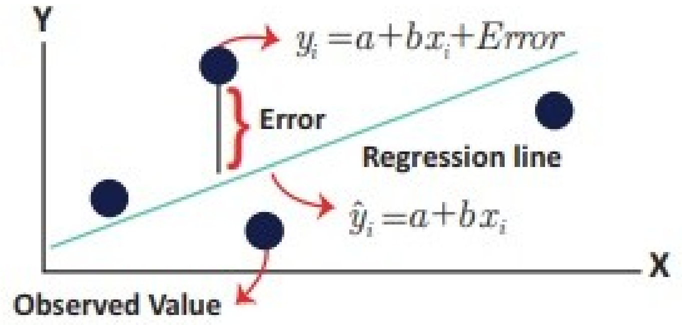


ID	Height (cm)	FEV1 (liters)
s1	164.0	3.54
s2	167.0	3.54
s3	170.4	3.19
s4	171.2	2.85
s5	171.2	3.42
s6	171.3	3.20
s7	172.0	3.60
s8	172.0	3.78
s9	174.0	4.32
s10	176.0	3.75
s11	177.0	3.09

The lest-square line

<u>Question</u>: which is the best line fitting the data?

- The one that minimizes *errors*
- Errors in terms *squared* deviation of points from the regression line



Method of the least-squares

→ find *a* and *b* that minimize:

$$\sum_{i=1}^{n} \left(y_i - (a + b \cdot x_i)\right)^2$$

We have analytical solutions...

Evaluating the regression equation

We are summarizing patterns of the data:

- It is inevitable that **assumptions** have to be made
- These assumption can be evaluated (eg. whether predictor have reasonably linear effect)
- Testing underlying assumption is especially important if specific claims are made on the **effect of the predictor**

Evaluating the regression equation

Inferential prospective:

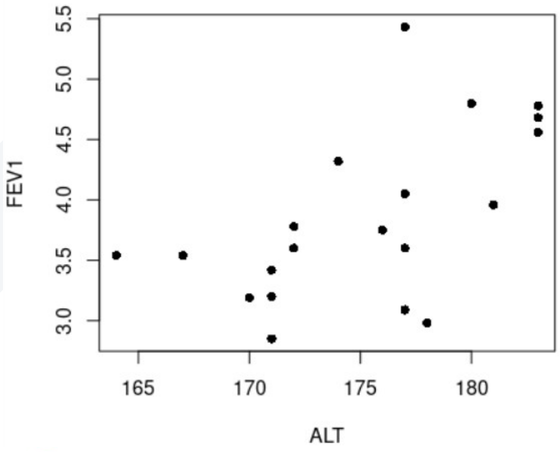
- *Y*, *X* and *E* are random variables
- *b* (**regression coefficient**) estimate: how to deal with uncertainty?
- Model fit: how to measure? When the model should be accepted?

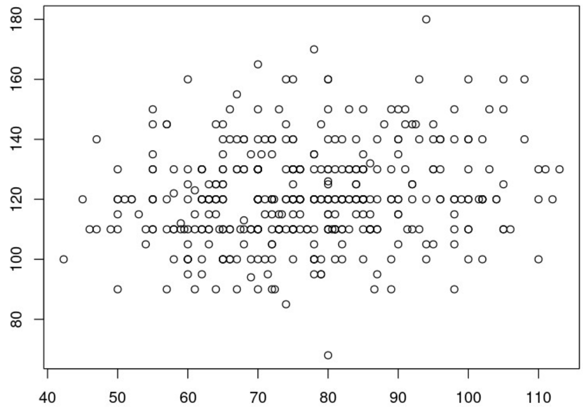
Main assumptions:

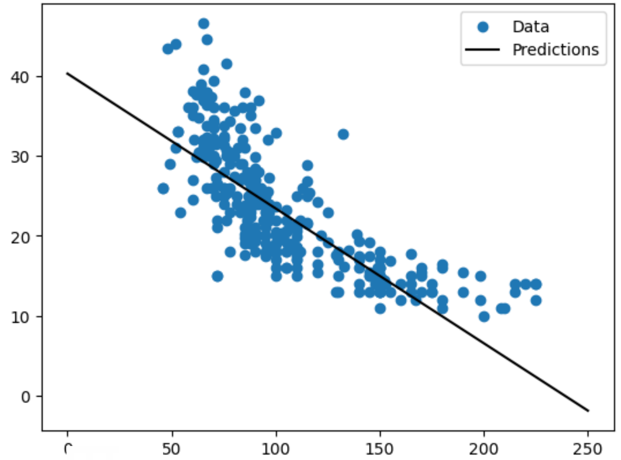
1. Linearity
2. Error term is normally distributed and has constant variance

Assumptions

1. **Linearity**: the relationship between *X* and *Y* can be expressed in a linear way







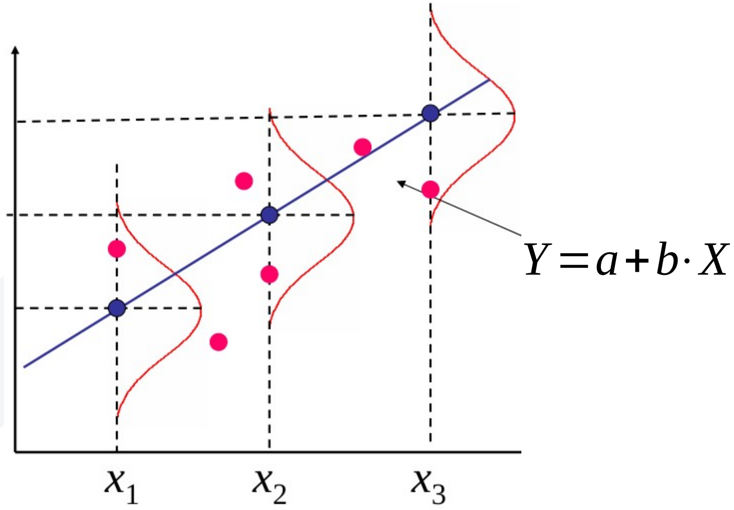
✔ Shows some linearity

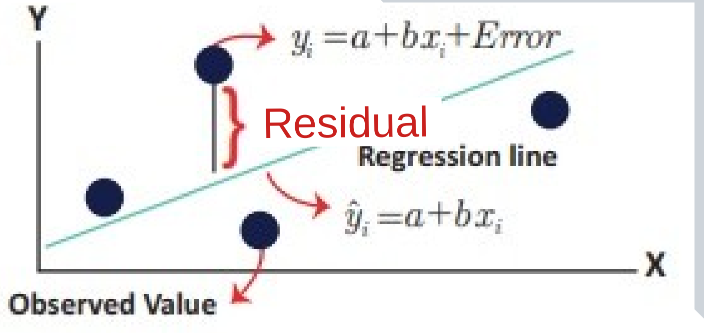
✔ Does not shows non-linearity

✘ Shows non-linearity

Assumptions

2. Error term: analysis of residuals



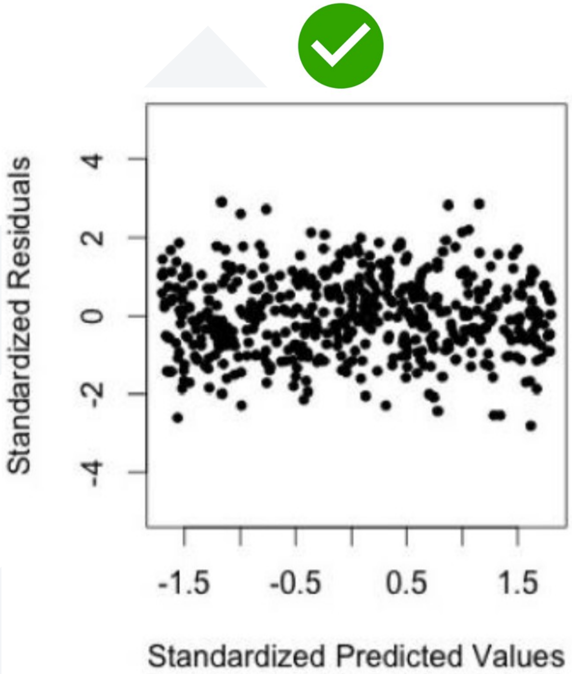


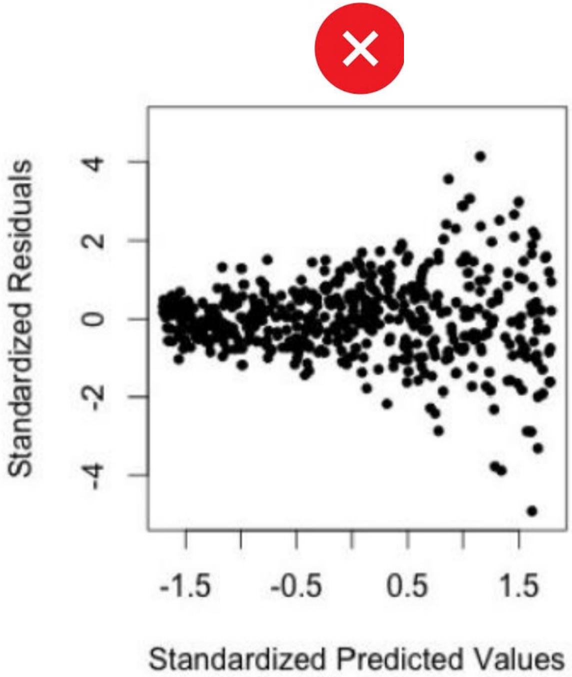
- To check **normality**: histogram, q-q plot
- To check **homoscedasticity**: plot residuals *vs* predicted values

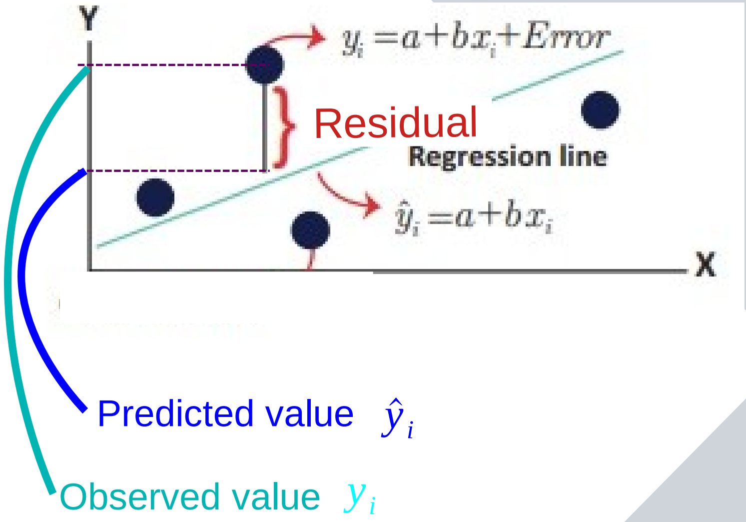
<u>Why</u>? the probability distribution of *b* depends on the distribution of the error term

Assumptions

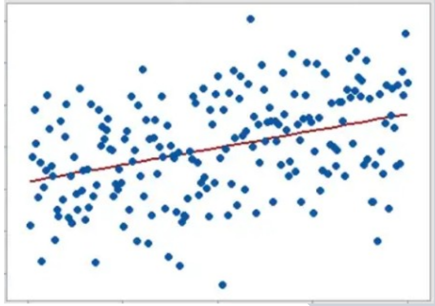
2. Error term: analysis of residuals

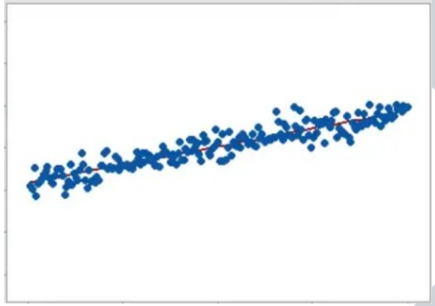






Goodness-of-fit





- $SS_{regression} = \sum (\hat{y}_i - \bar{y}_i)^2$: measures how values differ from the grand mean
- $SS_{residual} = \sum (y_i - \hat{y}_i)^2$: measures the error between predicted and observed values

We can define the **coefficient of determination**:

$$r^2 = \frac{SS_{regression}}{SS_{regression} + SS_{residual}}$$

***F*-test** can be performed to obtain the overall significance

It ranges between 0 and 1

Can be in interpreted as the proportion of **variance explained** by the dependent variable

Inference on the regression coefficient

- **Hypothesis testing**
  - $H_0: b=0$ this signifies no "relationship" or "effect"
  - Use of *t*-test

$$\frac{b}{SE(b)} \approx t_{n-2}$$

- **Confidence interval** for *b*:
  - $b \pm t^{\circ} \cdot SE(b)$

Model Coefficients - FEV1

			95% Confidence Interval			
Predictor	**Estimate**	**SE**	**Lower**	**Upper**	**t**	**p**
Intercept	-3.16	0.83	-4.80	-1.51	-3.79	0.0002
height	3.83	0.51	2.83	4.83	7.56	< .0001

Multivariable linear regression

A response variable is modelled against a linear combination of **two or more** simultaneously predictor **variables**:

$$Y = a + b_1 X_1 + \ldots + b_k X_k + E$$

- To explore the relationship between a response variables and two or more independent variables (or covariates“, “predictors”) appraised **simultaneously**
- To estimate the independent impact of a given covariate on the dependent variable, by **adjusting** for the contributions of all the other covariates

Multivariable linear regression

- **<u>Example</u>**: Effects on blood pressure (Y) of weight (X_1) and smoking (X_2) expressed as number of cigarettes per day)

$$Y = 37 + 0.01 \cdot weight + 0.5 \cdot cigarettes + E$$

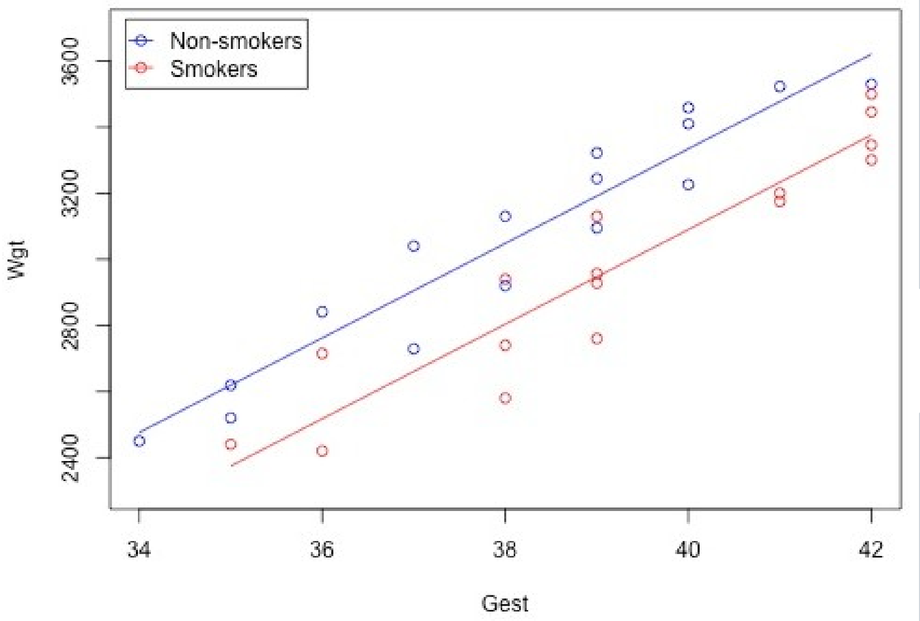
- b_i are **partial regression coefficients**: change of Y for 1 unit change of X_i and all the others $X_{j,j \neq i}$ remain constant
- 0.01 → average increase of y across subjects when weight is increased by 1 unit. if cigarette smoking is unchanged

Categorical predictors

Example: Effects on birth weight (Y) of length of gestation and smoking status (**yes/no**)

$$Y = -2390 + 143 \cdot gest - 244 \cdot smoker + E$$

- -244 : for smokers, on average, birth weight is reduced by 244g



If one of the predictors X_i is **binary**, b_i estimates the mean difference in Y for X_i=1 compared to X_i=0 → affects only the **intercept**

Categorical predictors

Model Term	Coefficient	Std. Error	t	Sig.	95% Confidence Interval Lower	Upper
Intercept	-1.672	.4705	-3.553	.000	-2.596	-.747
IL6	-.054	.0104	-5.146	.000	-.074	-.033
CRP	-.020	.0095	-2.131	.033	-.039	-.002
LengthofStay	-.115	.0358	-3.204	.001	-.185	-.045
CancerStage=IV	-2.210	.1537	-14.374	.000	-2.511	-1.908
CancerStage=III	-.947	.1028	-9.207	.000	-1.148	-.745
CancerStage=II	-.390	.0739	-5.285	.000	-.535	-.246
CancerStage=I	0[b]	.	.	.	.	.
Experience	.105	.0231	4.535	.000	.059	.150

Here, *CancerStage* has 4 groups

Effect of *stage IV* vs reference group

Stage I is the reference group

If one of the predictors X_i is **categorical**, with more than two groups, the comparison is performed by setting a **reference group** (thus we fall in the previous binary case)

Multivariable linear regression

(Obstet Gynecol 2013;121:46–50)

Correlation Between Birth Weight and Maternal Body Composition

Etaoin Kent, MRCOG, MRCPI, Vicky O'Dwyer, MRCPI, Chro Fattah, MD, Nadine Farah, MD,

Table 3. Multivariate Regression Analysis of Predictors of Birth Weight

Variable	Regression Coefficient (95% CI)	P
Gestational age at delivery (wk)	143.0 (129.6–156.4)	<.001
Fat-free mass	19.8 (17.0–22.7)	<.001
Smoking	−219.0 (−248.0 to 170.0)	<.001
Parity	124.7 (90.4–159.0)	<.001
Age (y)	3.3 (0.3–6.3)	.032
Fat mass	0.7 (−1.9 to 3.3)	.621

CI, confidence interval.
R^2=0.245.
Dependent variable: birth weight.
Independent variables: age, parity, gestational age at delivery, smoking, fat mass, and fat-free mass.

For one more gestational week,on average the weight increase is 143.0g

Being smoker, on average decreases the weight by 219.0g

Assumptions

1. Linearity
2. Error term is normally distributed and has constant variance
3. **No multicollinearity**: a predictor variable must not be correlated to other predictor variables (|r|>0.8)

Correlation: BP, Age, Weight, BSA, Dur, Pulse, Stress

	BP	Age	Weight	BSA	Dur	Pulse
Age	0.659					
Weight	0.950	0.407				
BSA	0.866	0.378	0.875			
Dur	0.293	0.344	0.201	0.131		
Pulse	0.721	0.619	0.659	0.465	0.402	
Stress	0.164	0.368	0.034	0.018	0.312	0.506

Correlation matrix

Logistic regression

What if the outcome of interest *Y* is a binary variable?

- disease/no disease
- dead/alive

A case-control study on hormone therapy as a risk factor for breast cancer in Finland: intrauterine system carries a risk as well

IJC
International Journal of Cancer

Heli K. Lyytinen[1], Tadeusz Dyba[2], Olavi Ylikorkala[1] and Eero I. Pukkala[2,3]

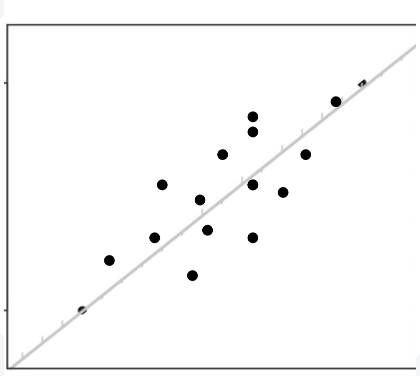
The purpose of this study was to evaluate the association between postmenopausal hormone therapy (HT) and the risk for breast cancer in recently postmenopausal Finnish women. All Finnish women with first invasive breast cancer diagnosed between the ages of 50 and 62 years during 1995–2007 (n = 9,956) were identified from the Finnish Cancer Registry. For each case, 3 controls of the same age were retrieved from the Finnish Population Register. The cases and controls were

Logistic regression

What if the outcome of interest *Y* is a binary variable?

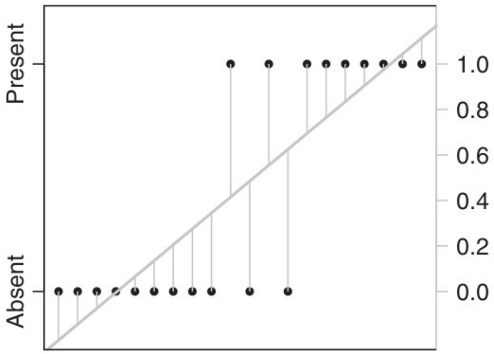
- disease/no disease
- dead/alive

Y



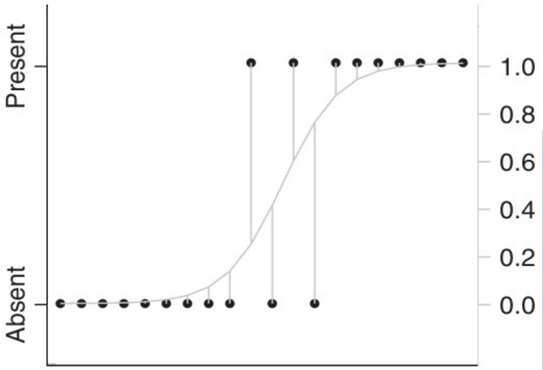
X

Y



X

Y



X

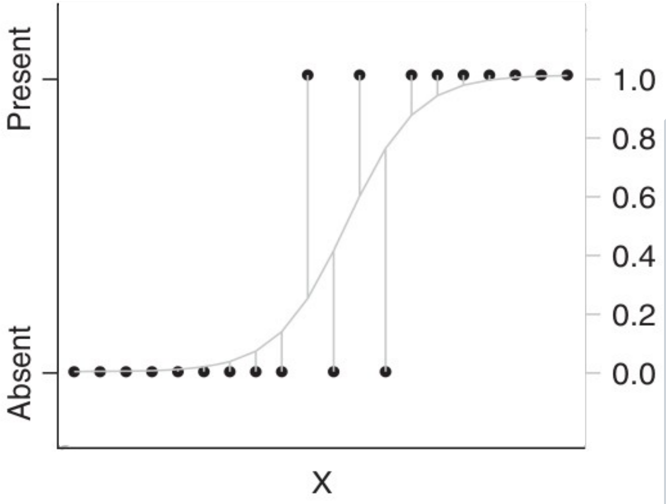
Logistic regression

It is aimed to model the effects of multiple predictors on a **binary response variable**

→ Y takes values 0 or 1 (disease *no* or *yes*)

Let's denote $P = E(Y) = P(Y=1)$

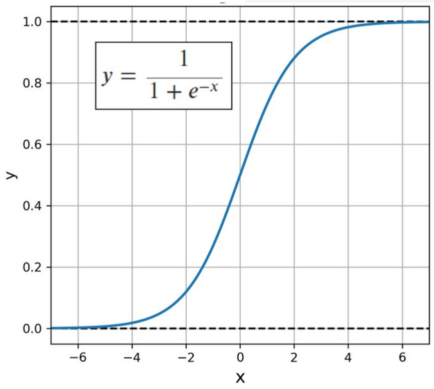
We can use a **non-linear** function to ***link*** response and linear combination of predictors:

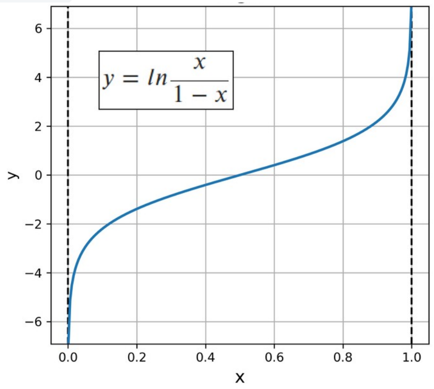


$$f(x) = \frac{1}{1+\exp(-x)}$$

Logistic regression

Sigmoid function





Logit function

$$P = P(Y=1)$$

$$P_i = \frac{1}{1+\exp(-(a+b\,x_i))} \qquad \log\left(\frac{P_i}{1-P_i}\right) = a + b\,x_i$$

Maximum Likelihood Estimation

After the logit transformation, the right side of the equation is linear

Logistic regression

Example: one continuous predictor

$$\log\left(\frac{P}{1-P}\right)=a+b\cdot BMI+E$$

P = P(diabetes)

What happens for one unit change in *BMI*?

$$\log\left(\frac{P'}{1-P'}\right)=a+bx$$

$$\log\left(\frac{P''}{1-P''}\right)=a+b(x+1)$$

$$e^b=\frac{\frac{P''}{1-P''}}{\frac{P'}{1-P'}}=OR$$

It's the **OR** obtained by increasing BMI of one unit

With respect to *b*, it's the **log odds ratio**: *b=log(OR)*

Odds Ratio (OR)

$$odds=\frac{P}{1-P}$$

$$odds\ ratio=\frac{odds\ of\ group\,1}{odds\ of\ group\,2}$$

$$odds\ ratio=\frac{P_1/(1-P_1)}{P_2/(1-P_2)}$$

Logistic regression

- P=P(heart disease)
- Predictors: age, weight, gender, VO2max

Odds Ratio (OR)

OR>1 increased odds for disease

OR=1 no change odds

OR<1 decreased odds for disease

To obtain the OR, we have to exp(*b*)

b is the logOR

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	age	.085	.028	9.132	1	.003	1.089	1.030	1.151
	weight	.006	.022	.065	1	.799	1.006	.962	1.051
	gender(1)	1.950	.842	5.356	1	.021	7.026	1.348	36.625
	VO2max	-.099	.048	4.266	1	.039	.906	.824	.995
	Constant	-1.676	3.336	.253	1	.615	.187		

a. Variable(s) entered on step 1: age, weight, gender, VO2max.

For a 1 year increase in age, the estimated OR is 1.089 → the risk (in *odds*) for heart disease is increased by 8.9%

Different from **probability**!

Logistic regression

Assumptions:

- The outcome is a **binary** variable
- There is a **linear relationship** between the **logit** of the outcome and each predictor variables
- Absence of **multicollinearity** among predictors
- There are no **influential values** (extreme values or outliers) in the continuous predictors

Check the residuals!

Logistic regression

A case-control study on hormone therapy as a risk factor for breast cancer in Finland: intrauterine system carries a risk as well

IJC
International Journal of Cancer

Heli K. Lyytinen[1], Tadeusz Dyba[2], Olavi Ylikorkala[1] and Eero I. Pukkala[2,3]

Table 3. Relative risk of invasive breast cancer among postmenopausal women using hormone therapy

Therapy	Cases	Controls	OR[1]	95% CI	*p*
No user[2]	5,473	17,956	1.00	(Reference)	
Estradiol-only therapy	991	3,300	1.01	0.93–1.09	0.88
Progestagen-only therapy	138	476	0.97	0.80–1.17	0.73
LNG-IUS[3]	329	708	1.53	1.33–1.75	0.001
Estradiol-progestagen therapy	1,731	4,243	1.36	1.27–1.46	0.001
Estradiol plus LNG-IUS	287	473	2.07	1.78–2.41	0.001
Mixed therapy[4]	927	2,534	1.22	1.12–1.33	0.001
Tibolone	80	178	1.36	1.15–1.96	0.003

[1]Adjusted with age, parity, age at first birth and health care district. [2]Had bought HT never or for less than 6 months. [3]Levonorgestrel releasing intrauterine system. [4]Mixture of estradiol-only, progestagen-only, estradiol-progestagen therapy, or tibolone.

"A multivariate conditional logistic regression model was used to estimate, by means of the odds ratio (OR), the relative risk for breast cancer associated with each category of HT use"

Although not shown, multiple predictors were included in the model

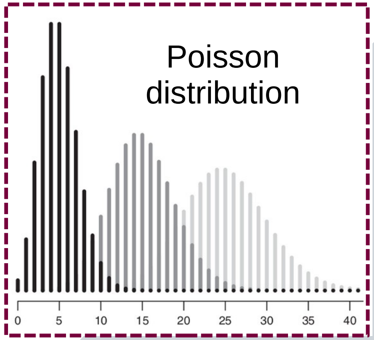
Generalized Linear Models

GLM provide a set of recognized procedures for relating response variables to a **linear combination** of one or more predictors:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

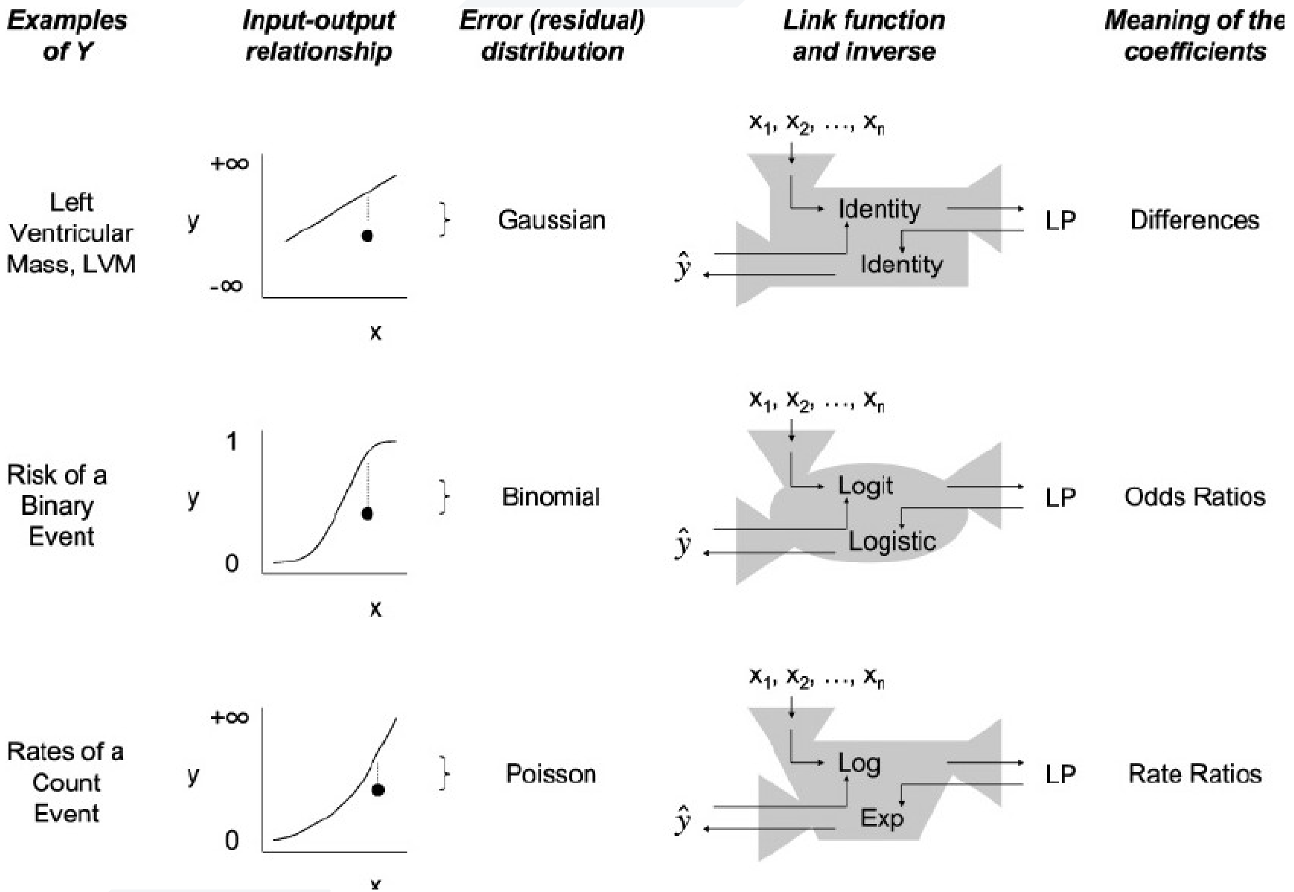
Where *g(μ)* represents the **link function**

Model	Response variable	Predictor variable(s)	Residual distribution	Link
Linear regression[a]	Continuous	Continuous/ Categorical	Gaussian (normal)	Identity $g(\mu) = \mu$
Logistic regression	Binary	Continuous/ Categorical	Binomial	Logit $g(\mu) = log_e \frac{\mu}{1-\mu}$
Log-linear models	Counts	Categorical	Poisson	Log $g(\mu) = log_e \mu$



For count count data

Generalized Linear Models



When working with GLM the **interpretation** of the predictor effects becomes more challenging

Ravani. *Clinical research of kidney diseases IV: standard regression models*. Nephrol Dial Transplant (2008)

Which predictors?

Ideally, every epidemiological study would be designed with attention given to a small set of risk factors, and a further set of possible confounding variables identified *a priori*

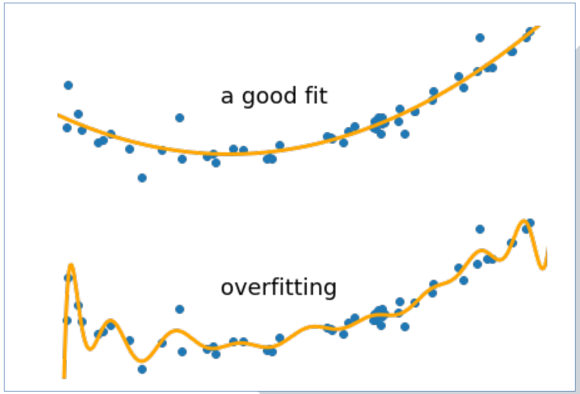
The exact nature of risk factors could be unknown in the study design phase (limited prior knowledge) and many possible candidate exposure variables (including *proxies*) are measured $\rightarrow$ strategies for **model building**

Sample size

Missing data

If a subject presents a missing value in one of the predictors, he will be completely removed from the analysis

- When estimating regression models an adequate *effective* **sample size** must be ensured
- If the fitted model is too complex (too many predictors for the amount of information in the data), the goodness of fit of the model will be *exaggerated* and future observed values will not agree with the predicted values (**overfitting,** lack of **generalization**)



Sample size

Rule-of-thumb: a fitted regression model is likely to be reliable when the number of predictors *p* is less than *m*/10 or *m*/20, where *m* is the limiting sample size

Type of Response Variable	Limiting Sample Size m
Continuous	n (total sample size)
Binary	$\min(n_1, n_2)$ [h]
Ordinal (k categories)	$n - \frac{1}{n^2}\sum_{i=1}^{k} n_i^3$ [i]
Failure (survival) time	number of failures [j]

Linear regression
For 3 predictors we need more than 3*10=30 individuals

Logistic regression
Assuming that cases is the rarer category, for 3 predictors we need more than 3*10=30 cases (and >30 controls → >60 individuals)

Survival model
For 3 predictors we need more than 3*10=30 failure events (eg. deaths)

Sample size

An appropriate **study design** is essential:

- Number of predictors: we must pursue **parsimony** in model specification
- If there are known associated predictors (eg. known **risk factors**, **confounders**) to our response variable, these must be included in the model and this will increase the complexity
  - Adequate sample size!
- **subject-matter knowledge** should guide multivariable model-building

Variable selection

- Variable selection is used when we face with many **potential predictors** but we don’t have the necessary prior knowledge to prespecify the *important* ones to be included
- There is a rich set of techniques that **algorithmically** search through subsets of the predictors in attempting to choose a model that both fits the data well and also does not include many unnecessary variables
- The choice of the approach depends on the **aim** of model building

Different scientific aims

Descriptive modelling

Aim: to capture the data structure

Characteristics:
- Interpretability
- Transportability
- Parsimony

Predictive modelling

Aim: to predict new or future observations

Characteristics:
- Accuracy
- Complexity allowed

Explanatory modelling

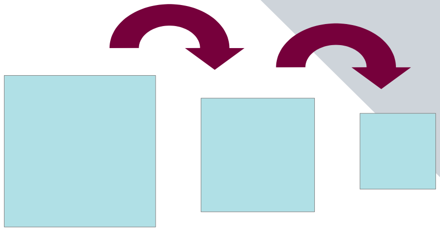
Aim: to test causal theory

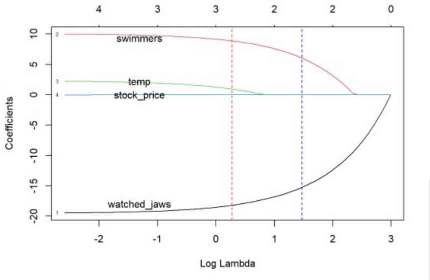
Characteristics:
- Starts from theoretical constructs
- Conclusion often converted into *policy* recommendations

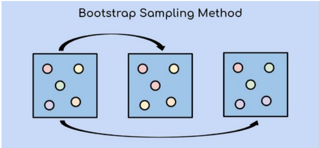
Variable selection

The choice depends on the aim of the model!

- Based on subject matter knowledge
- Stepwise selection: the fit of many variable combinations is compared using Information Criteria
  - Akaike's (AIC): preferable for *predictive* models
  - Bayesian (BIC): preferable for *descriptive* models
- LASSO penalization
- Resampling-based procedures
- And more… (often rooted in *machine learning* field)







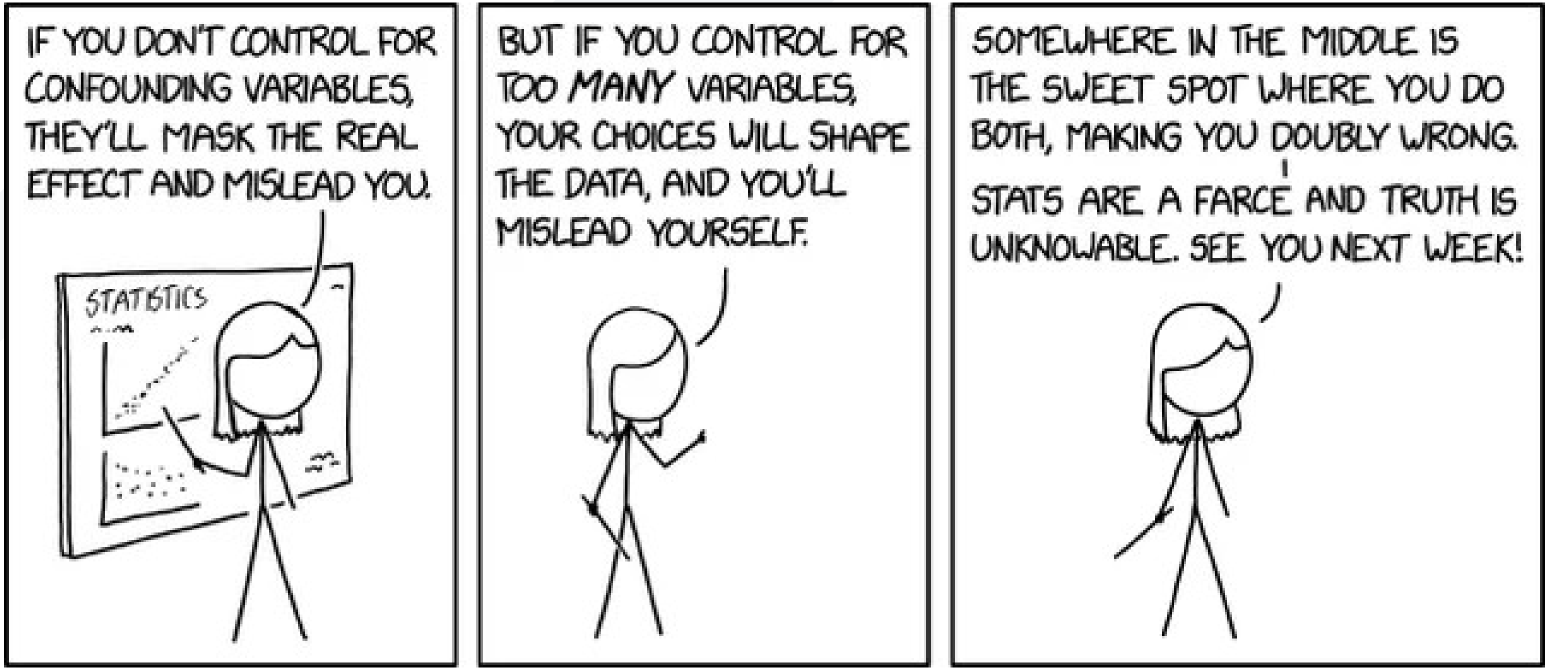
Variable selection

What constitutes a good variable for prediction and a good variable for significance depend on different properties of the underlying distributions:

- Significant variables: may be associated with the outcome simply for a small group of individuals, thereby leading to poor prediction
- Predictive variables: may be influential for the outcome but not necessarily appear highly significant (for a particular hypothesis)

Statistical significance does not imply practical importance, and conversely

Variable selection



https://xkcd.com/2560/

Validation of model predictivity

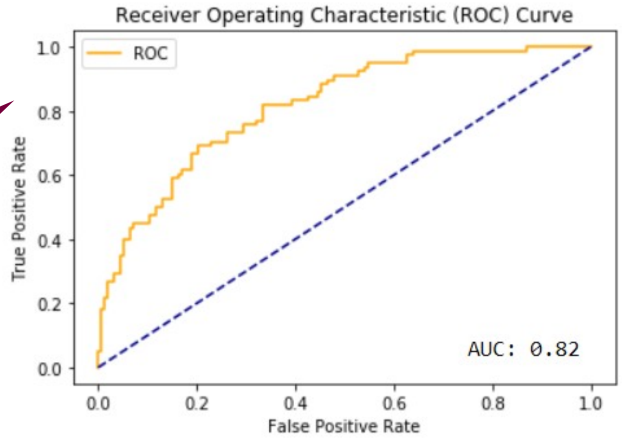
We would like to ascertain whether predicted values from the model are likely to accurately predict responses on **future subjects** or subjects not used to develop our model → **validation**

<u>Example</u>: logistic model for diabetes

- The model returns a value P_i for each subject

$$P_i = \frac{\exp(a + b\,x_i)}{1 + \exp(a + b\,x_i)}$$

Area Under the ROC Curve (**AUC**)



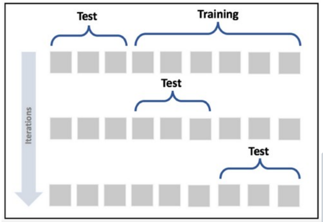
- Can be used to classify diabetic *vs* non-diabetics?

Validation of model predictivity

If predictivity (eg. discrimination ability) is measured on the data used to derive the model, we will get **overoptimistic results**

Two major ways of model validation:

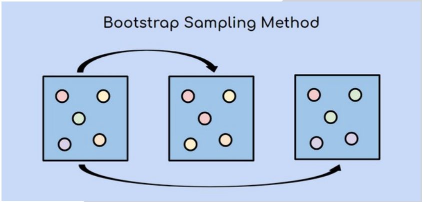
Requiring more data



- Use of a separate validation cohort (***external***)

Higher computational cost

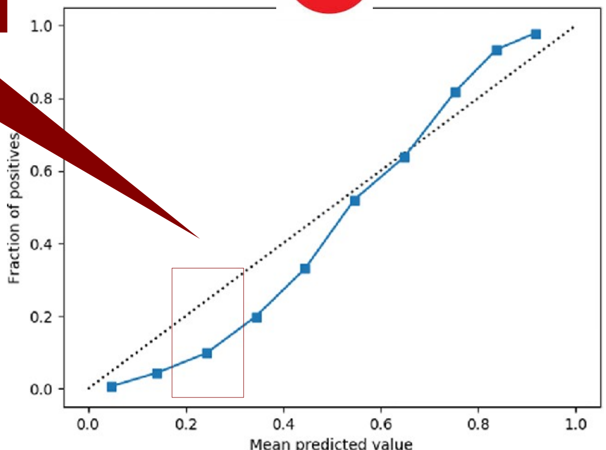
- **Resampling** methods (***internal***)
  - <u>Cross-validation</u>: reserving a subsample to test the model
  - <u>Bootstrap</u>: mimic the process of obtaining new datasets



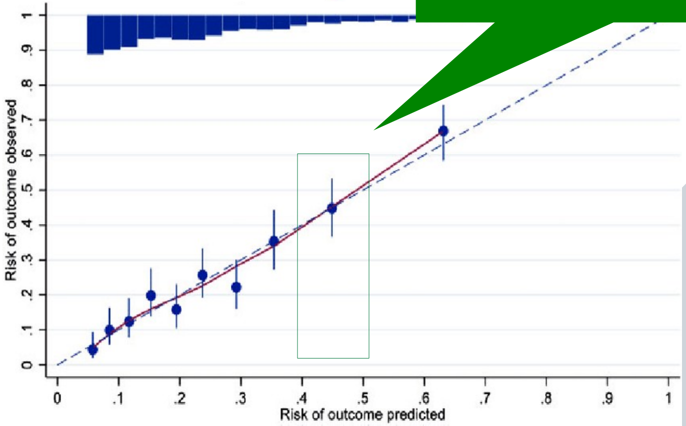
Validation of model calibration

- Discrimination is important, but are the risk estimates ***reliable***?
- **Calibration plot**: observed responses against predicted responses

Here risk estimate are systematically too low



Here the proportion of individuals with risk between 0.4-0.5 is between 40-50%



Is the model useful?

Many predictive models are never used...

- It was not deemed relevant to make predictions in the setting envisioned by the authors
- Potential users did not trust the relationships, weights or variables used to make the predictions
- The variables necessary to make the predictions were not routinely available

References

- Harrell, Frank. (2010). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. 10.1007/978-1-4757-3462-1.

- E. Suárez, Erick & Pérez, Cynthia & Rivera, Roberto & Martínez, Melissa. (2017). *Applications of Regression Models in Epidemiology*. 245-250. 10.1002/9781119212515.index.

- E. W. Steyerberg. *Clinical prediction models*. Springer Cham. 2019

- G. Shmueli. *To explain or to predict?* Statistical Science 2010, Vol. 25, No. 3, 289–310

- W. Sauerbrei et al. *State of the art in selection of variables and functional forms in multivariable analysis - outstanding issues*. Diagn Progn Res. 2020;4(1):1–18