UNIVERSITÄTS KLINIKUM FREIBURG

Lecture 1
Introduction to regression models and issues in analysis

Willi Sauerbrei, Edwin Kipruto

Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center -University of Freiburg, Germany

WS 2025/26

After this lecture, you will be able to ...

- Know that regression models are the key tool for many analyses
- Outcomes may be different, but regression part is very similar
- Significance tests can be represented in a regression model
- Many factors can be investigated simultaneously in a regression model
- Understand three important models
  - Linear regression model
  - Logistic regression model
  - Cox regression model for survival data
- Distinguish between modelling aims
  - Description
  - Prediction
  - Explanation/causal effects

To Describe, to Predict or to Explain?

- **Descriptive models**
  - Capture the data structure parsimoniously
    - Which variables are associated with the outcome and how? ← VARIABLE SELECTION
    - Smoothing/functional forms: efficient estimation of expected values ← FUNCTIONAL FORM ESTIMATION
  - **Prediction models**
    - Interest in accurate predictions for future application
  - **Explanatory (causal) models**
    - Interest in effect of an intervention on an individual's outcome

Often several modeling goals simultaneously:

- Transparent prediction models (D + P)
- Counterfactual prediction models (E + P)

- Shmueli, G. (2010). To explain or to predict?. *Statistical science*, 289-310.
- Carlin, J. B., & Moreno-Betancur, M. (2025). On the uses and abuses of regression models: a call for reform of statistical practice and teaching. *Statistics in Medicine*, *44*(13-14), e10244.
- Sauerbrei, W., Ambrogi, F., de Bin, R., Boulesteix, A. L., Goetghebeur, E., & Huebner, M. (2025). Commentary: Regression Models—Efforts Are Required to Improve Statistical Practice and Teaching. *Statistics in Medicine*, *44*(13-14), e10341.

What is Regression?

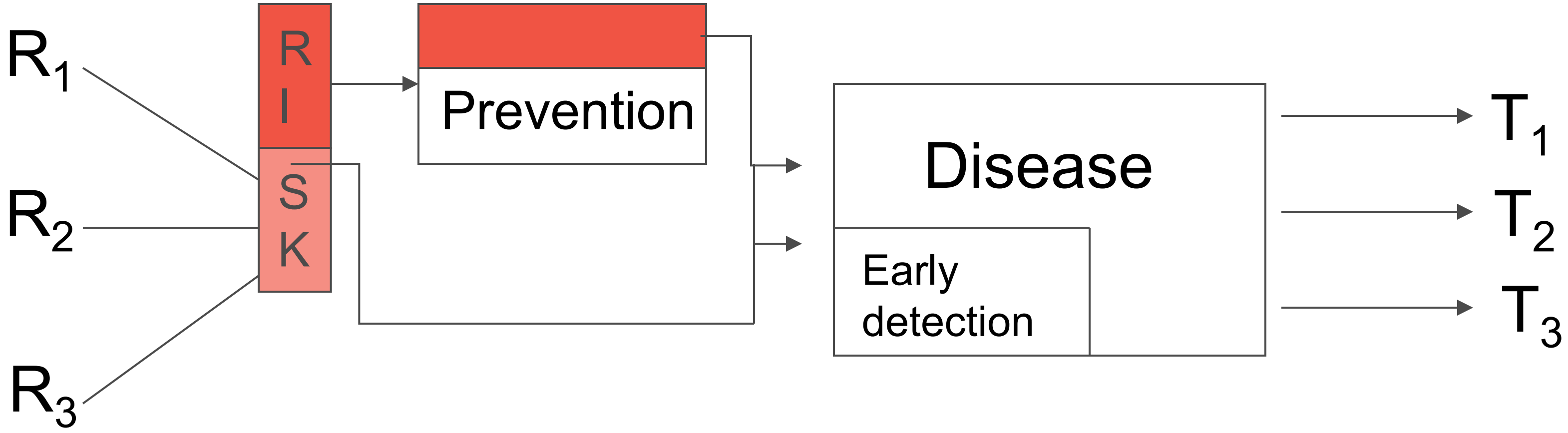
- Statistical method to investigate the association between a response variable Y and one or more explanatory variables X_1, ..., X_k
- Response variable Y may be continuous, binary or a survival time (partly censored).
- X_1,..., X_k may be risk factors, prognostic factors, diagnostic criteria etc.

Trials to gain clinical knowledge

Evaluation of risk factors

Diagnosis

Comparison of therapies



Epidemiologic Study Potential problems	Prevention study	Screening-, Diagnosis-study	Therapy studies (+ prognostic factors)
Relevant risk factor?	Unneeded intervention?	False pos/neg diagnosis	Unsuitable therapy

Gaining medical knowledge

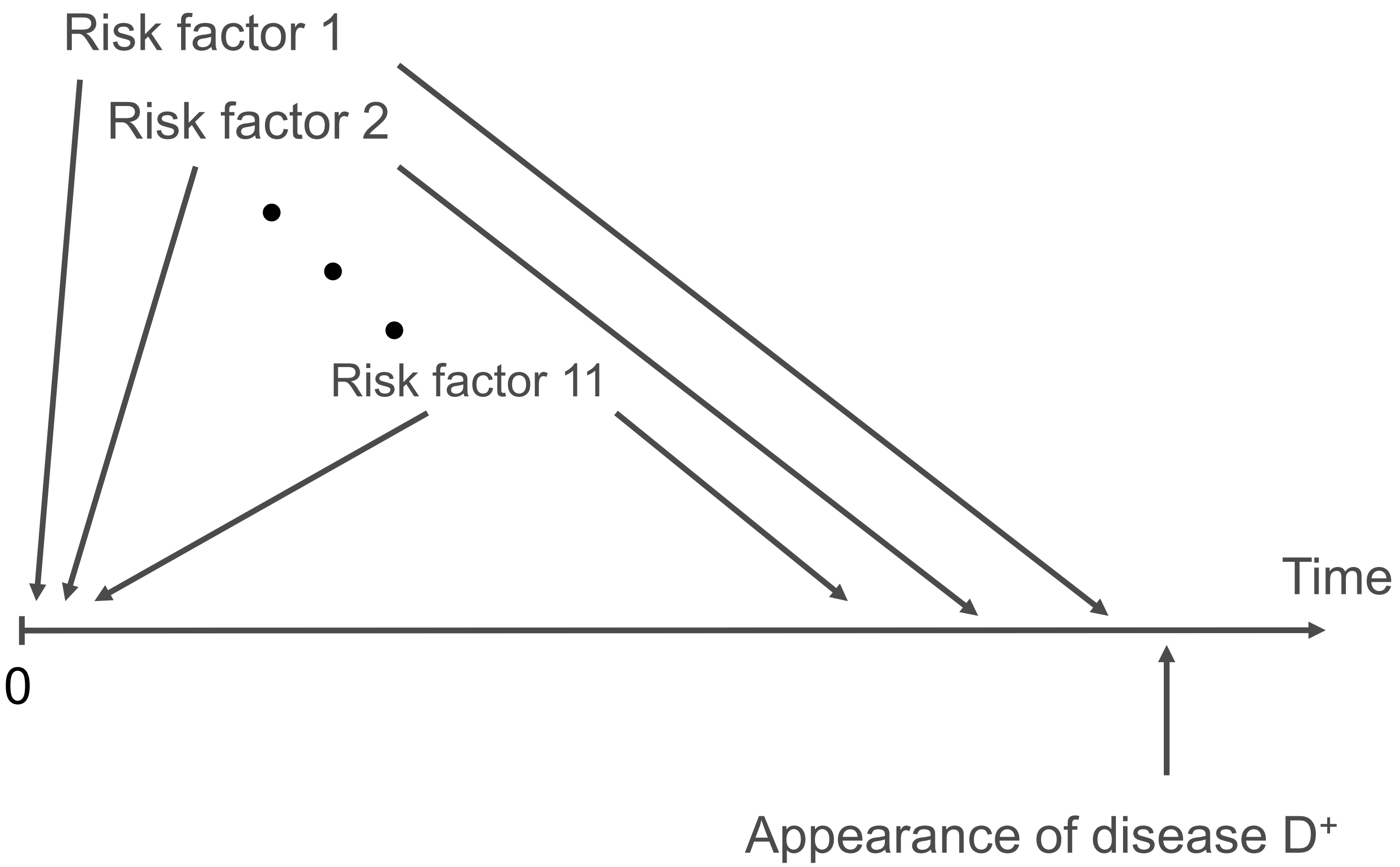
Important basics

- “Good” trials
- “Good” data
- Reasonable analysis, summary and interpretation of the data

However, it is no exception that there are…
… bad trials, bad analysis, wrong interpretation, false conclusions and

very bad reporting

Many trials investigate complex questions, e.g.: Risk factors can influence the development of a disease in complex ways



Usually several factors have an influence on the outcome.

Regression models are the **key tool** for the analysis of most of these studies. However, several alternatives (e.g. trees, neural nets) are available. Not considered here.

Outcome and regression model

Different types of regression models for different outcomes

A) Linear Regression: **Y continuous**

1. Y=Weight, X=Height
2. Y=FEV1, X_1=Weight, X_2=Height, X_3=Age

Simple linear regression: $Y=\beta_o + \beta_1 + e$

Multiple linear regression: $Y=\beta_o + \beta_1 X_1 + ... + \beta_k X_k + e$

B) Logistic Regression: **Y binary response**

1. Y=Deceased yes/no
2. Y=Case yes/no

$p=P(Y=1) \quad \in \quad [0,1]$

$odds=p/(1-p) \quad \in \quad [0,\infty[$

$\log odds=\textit{logit}=\log(p/(1-p)) \quad \in \quad]-\infty,+\infty[$

Specification of a meaningful association between Y and X via the logits

Model: $$logit(\mathrm{p}) = \log\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k$$

$$\left.\Leftrightarrow \mathrm{p} = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}\right\}$$ “logistic“ function

β_i is logistic regression coefficient (adjusted)

$\exp(\beta_i)$ = Odds Ratio (adjusted)

C) Survival data (later)

Principles of regression models

- Regression models ...
  - typically relate one or more covariates to one response
  - allow for various response types (continuous, binary, time-to-event), which require specific types of models
  - require explicit specification of the influence structure of the covariates on the response

- Simplest case: linear regression
  - Continuous response
  - Additive, linear influence of the covariates
  - Well understood: Most regression modeling techniques are extensions of linear regression

For ‚well behaved‘ data (sample size $N \gg$ number of explanatory variables, no extreme cases, no complex correlation structure between Xs etc.) statistical approaches for estimation, testing etc. are available. Often procedures and their properties are first developped in the linear regression model and then transferred to the other models. Sometimes the transfer creates new problems in LR and survival time models.

Linear regression

Estimation

In classic statistics we minimize the sum of squares

$$S^2 = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2 = \sum_{i=1}^{n} \left(e_i\right)^2$$

This is the sum of the squared deviations of the observed points from the regression line parallel to the y-axis!

The minimisation problem could be expressed in matrix notation as:

$$X^t X\beta = X^t y \leftrightarrow$$

$$\beta = (X^t X)^{-1} X^t y$$ if $(X^t X)$ has an inverse

Some terms

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

…is called a predicted or fitted value

$$\hat{e}_i = Y_i - \hat{Y}_i$$

…is called a (raw) residual

This is called classic linear regression, when we may assume, that the residuals are normally distributed:

$$e_i \sim N(0, \sigma)$$

Goodness of fit

- Part of the idea of regression is to see, how well the variation in an explanatory variable explains the variation in the dependent variable

- This can be formalized as a description of goodness of fit of the fitted model

$$r^2 = r_{xy}^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2},$$

Residual analysis

- The residuals $\hat{e}_i = Y_i - \hat{Y}_i$ should be „sort of“ normally distributed

- They should be random with equal variance
  - Check this by looking at scatterplots of residuals vs predicted value or x
  - A „shape“ is bad
  - Wider or slimmer cloud at one end, toward the ends, or toward the middle is bad

More than one explanatory variable

- Extending linear regression to more than one explanatory variable is straight forward

- $Y_i = ß_0 + ß_1x_{1i} + ß_2x_{2i} + \ldots + ß_px_{pi}$

- The coefficients are now a vector of length p+1

$$\hat{\beta} = (\mathrm{X}^{\mathrm{t}}\mathrm{X})^{-1} X^t y \text{ if } (\mathrm{X}^{\mathrm{t}}\mathrm{X}) \text{ has an inverse}$$

- If e.g. x_1 and x_2 are somewhat correlated, then the estimate of $ß_1$ is different whether you include x_2 in the model or not

- Confounding!

Confounding

Estimate effect of X_1 on Y.

Wrong result if X_2 is ignored!

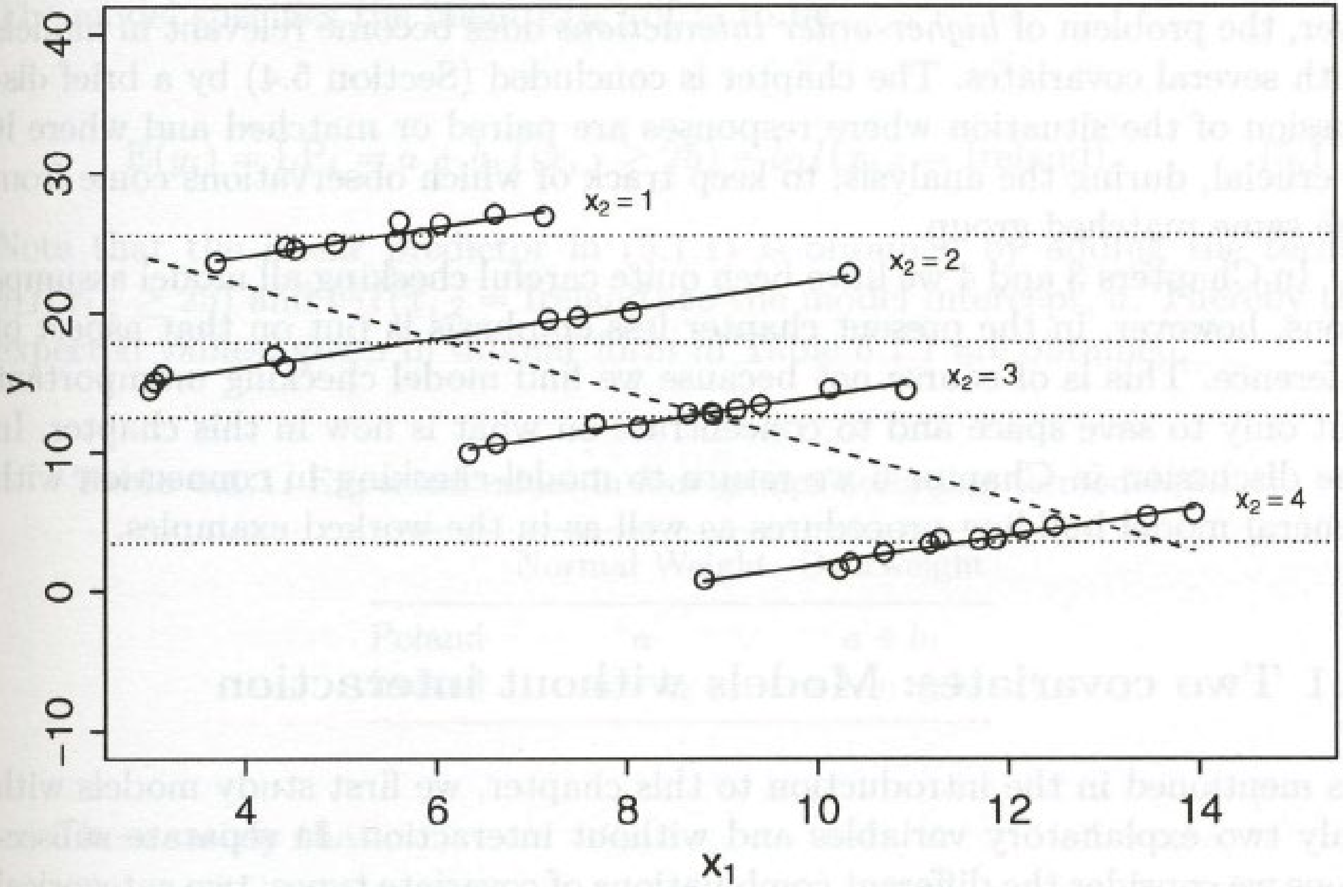


Fig. 5.0.1. Illustration of a confounding categorical variable x_2 when the effect of a quantitative variable x_1 on y is studied; see text.

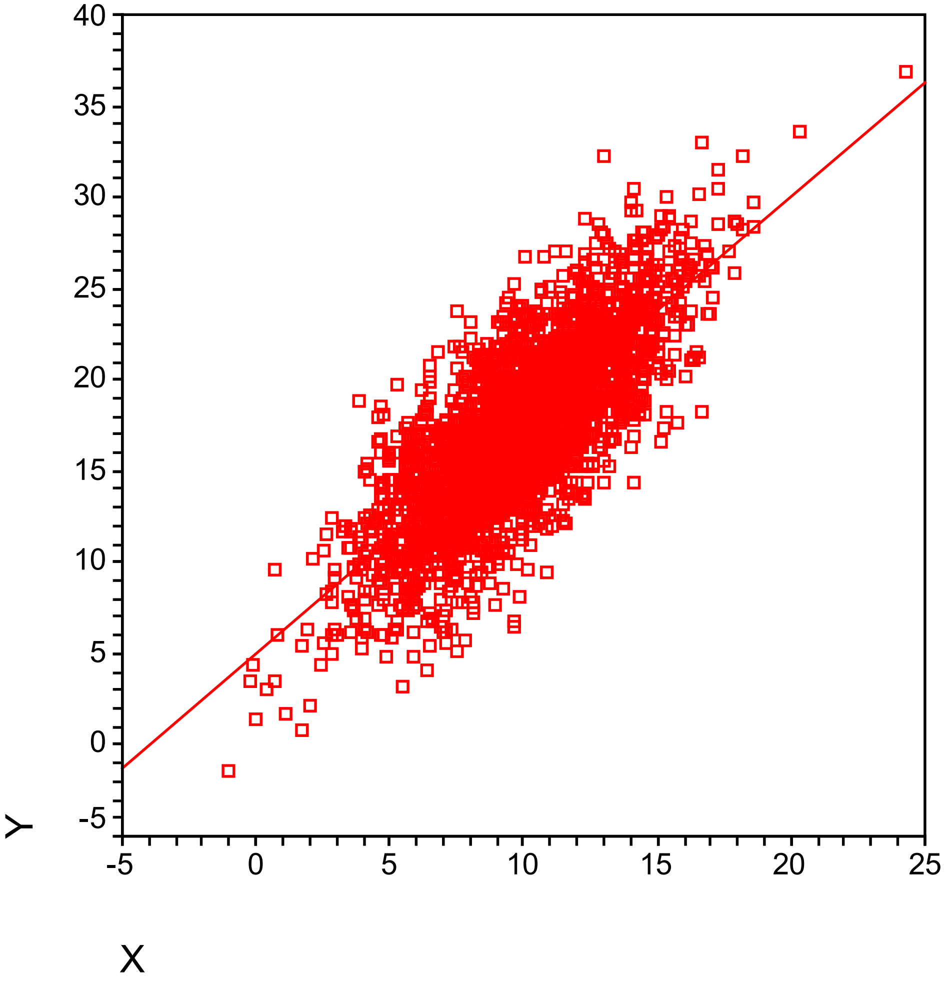
Andersen and Skovgaard, Springer 2010

R^2 Improvement

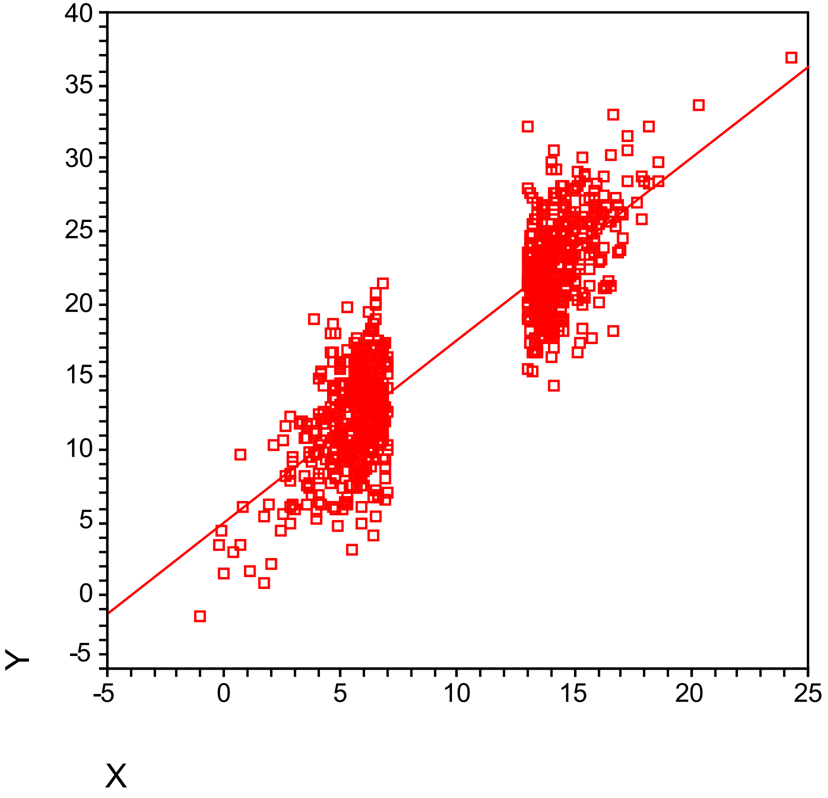
- Let $R^{2(p+1)}$ be the goodness of fit statistic for a model with p+1 explanatory variables
- Let $R^{2(p)}$ be the goodness of fit statistic for the nested model with p explanatory variables
- Then $R^{2(p+1)} > R^{2(p)}$
- BUT: If the improvement is very small, it is not worth „spending“ one more df
- Given a dataset, try to find the model with a good R^2 while at the same time being parsimonious about the number of explanatory variables!
    - Criteria: p-values, R^2 improvement, F-test for groups of variables (see ANOVA), relevance of explanatory variable

Interpretation of R^2

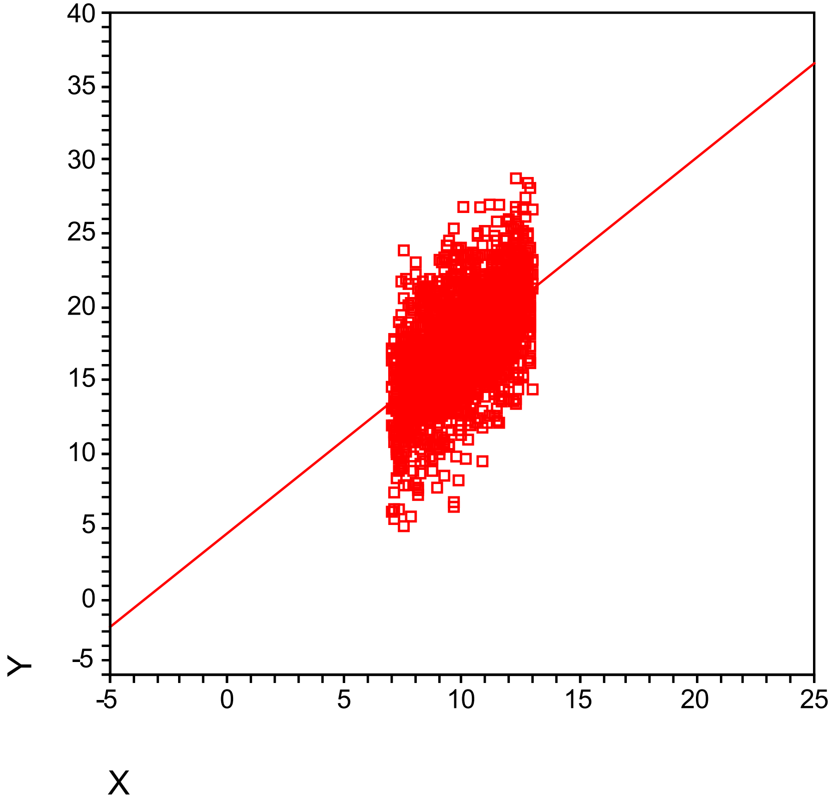
Overall



X ≤ 7 or X ≥ 13



7 < X < 13



Interpretation of R^2

True value		Estimate		
		overall	$x \leq 7$ or $x \geq 13$	$7 < x < 13$
n	∞	2500	768	1732
β_0	5.0	4.9	5.0	4.7
β_1	1.25	1.25	1.24	1.27
σ^2	9.0	9.05	9.25	8.95
R^2	0.61	0.60	0.79	0.32

Assessing the influence of a factor

1. Test (for one factor)

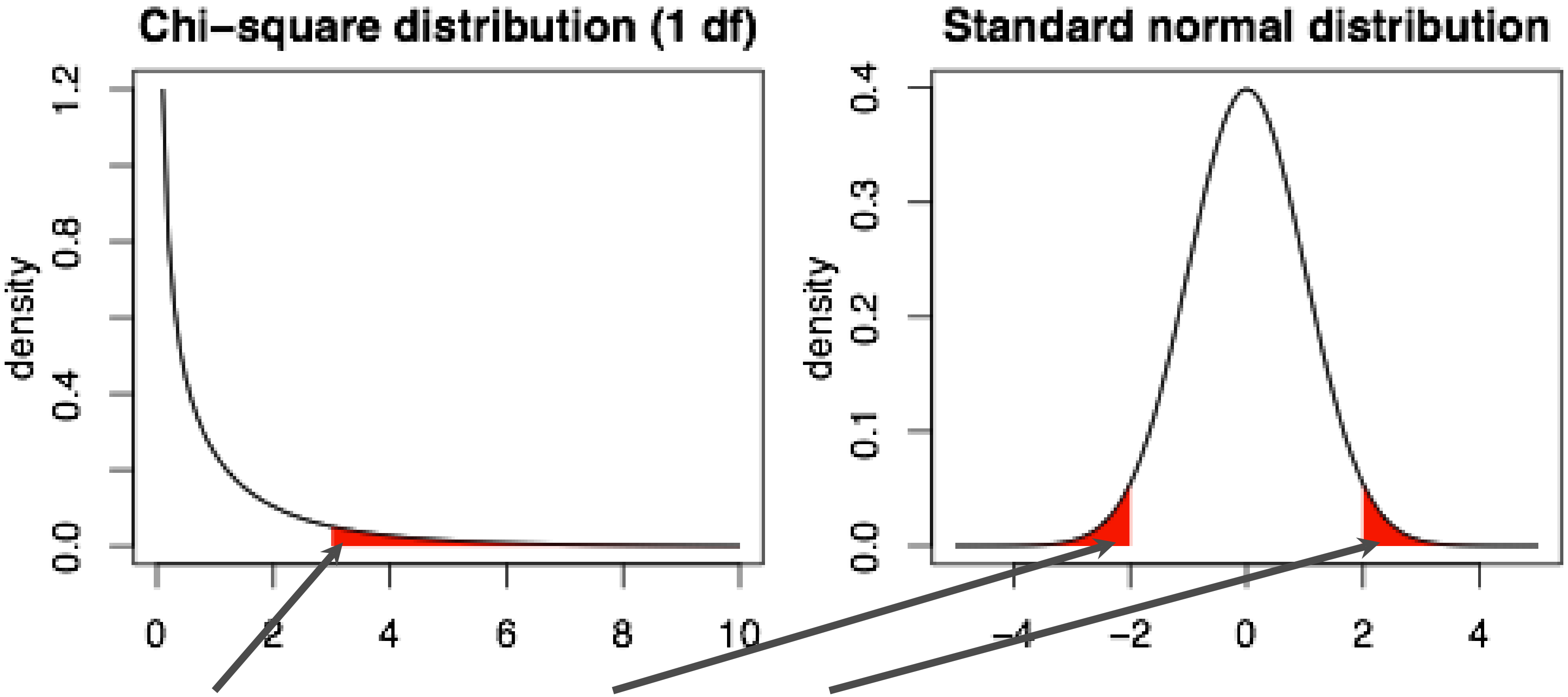
2. Regression models (one or more factors, much more flexible)

Principles of statistical testing

- Starting point: subject matter hypothesis about some population properties in the form "There is a ...", "There exists ...", "... is connected to ...", etc.
- Statistical version:
  - Null hypothesis (H_0): "There is no ..."
  - Alternative hypothesis (H_1): "There is a ..."
  - Ideally, we want to reject the null hypothesis on the basis of some data
- Find some test statistic that …
  - can be calculated from a sample of observations
  - has known distribution under the null hypothesis
- Obtain a sample of observations
- If the value of the test statistic provides enough evidence against H_0, reject the null hypothesis

Distributions and p-values

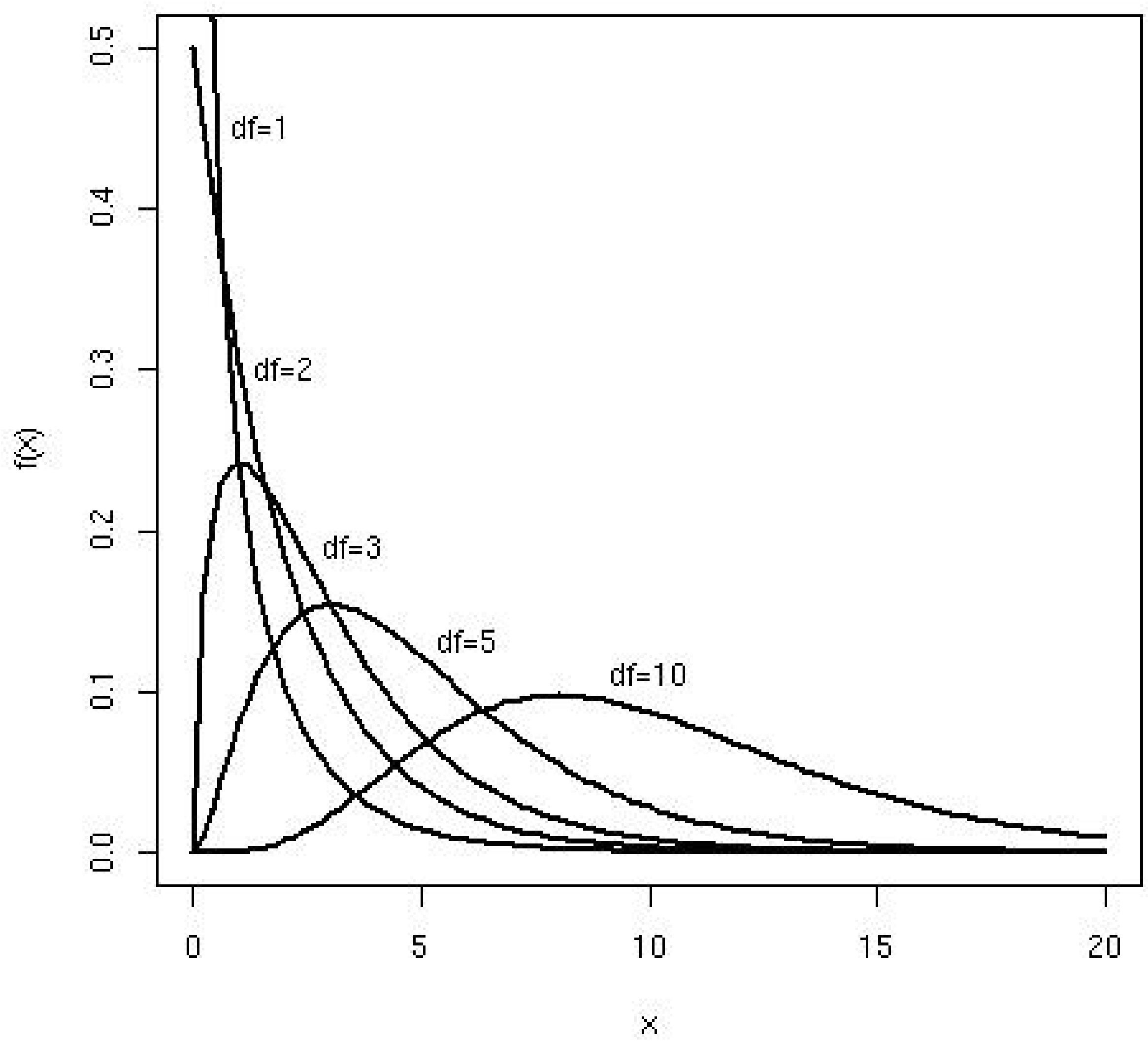
Two frequently encountered distributions:



Areas beyond the observed values provide p-values, which quantify the evidence against the null hypothesis

Convention: p ≤ 0.05 ⇒ "significant effect"

χ^2 (Chi-square) distribution with q degrees of freedom(df)



df	Crit. value	
q	α=5%	α=1%
1	3.84	6.63
2	5.99	9.21
3	7.82	11.34
5	11.07	15.09

Statistical Test

Decision matrix

	Actual situation	
Decision for	H_0 (no diffference)	H_1 (difference)
H_0 (no difference)	Correct decision $(1 - \alpha)$	False negative decision *Type 2 error* (β)
H_1 (difference)	False positive decision *Type 1 error* (α)	Correct decision $(1 - \beta)$

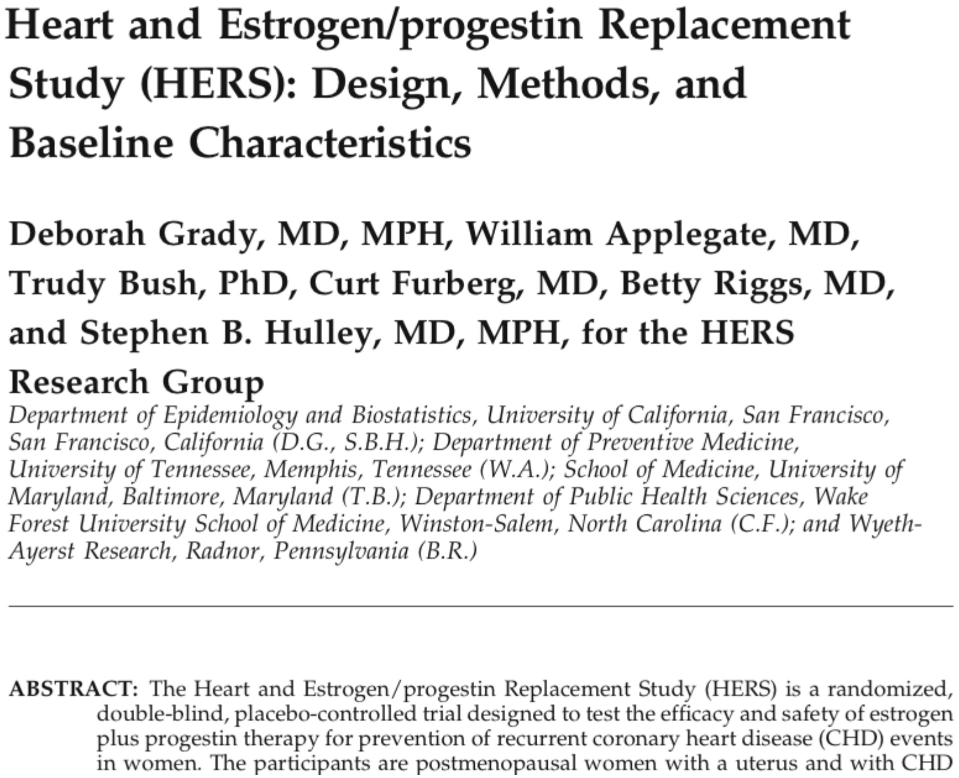
Requirement: The probability for

type 1 error: smaller than the **given value** α (signifiance level)

type 2 error (β**):** as small as possible.

Heart and Estrogen/progestin Replacement Study (HERS)

- Sample: observational data of 2028 women
- Outcome variable: blood glucose level
- Question: Influence of physical activity (“yes = 3 or more times per week”/ “no = less than 3 times per Woche”)



Heart and Estrogen/progestin Replacement Study (HERS): Design, Methods, and Baseline Characteristics

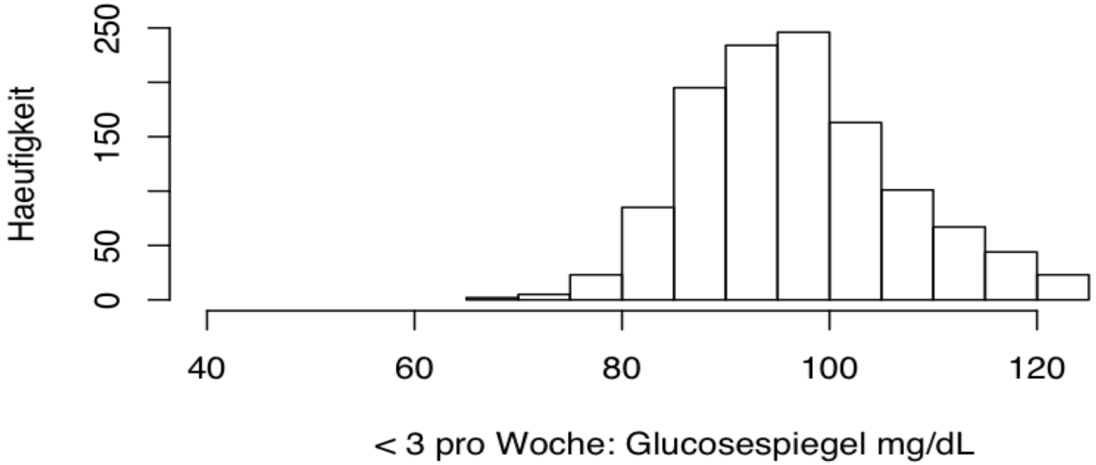
Deborah Grady, MD, MPH, William Applegate, MD, Trudy Bush, PhD, Curt Furberg, MD, Betty Riggs, MD, and Stephen B. Hulley, MD, MPH, for the HERS Research Group

Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California (D.G., S.B.H.); Department of Preventive Medicine, University of Tennessee, Memphis, Tennessee (W.A.); School of Medicine, University of Maryland, Baltimore, Maryland (T.B.); Department of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, North Carolina (C.F.); and Wyeth-Ayerst Research, Radnor, Pennsylvania (B.R.)

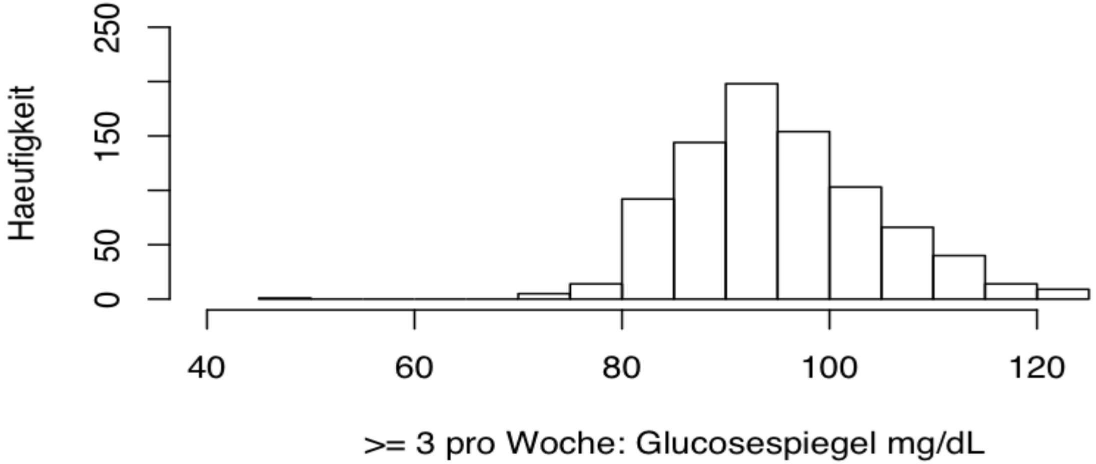
ABSTRACT: The Heart and Estrogen/progestin Replacement Study (HERS) is a randomized, double-blind, placebo-controlled trial designed to test the efficacy and safety of estrogen plus progestin therapy for prevention of recurrent coronary heart disease (CHD) events in women. The participants are postmenopausal women with a uterus and with CHD

Controlled Clinical Trials, 1998, 19(4):314-335

Histograms

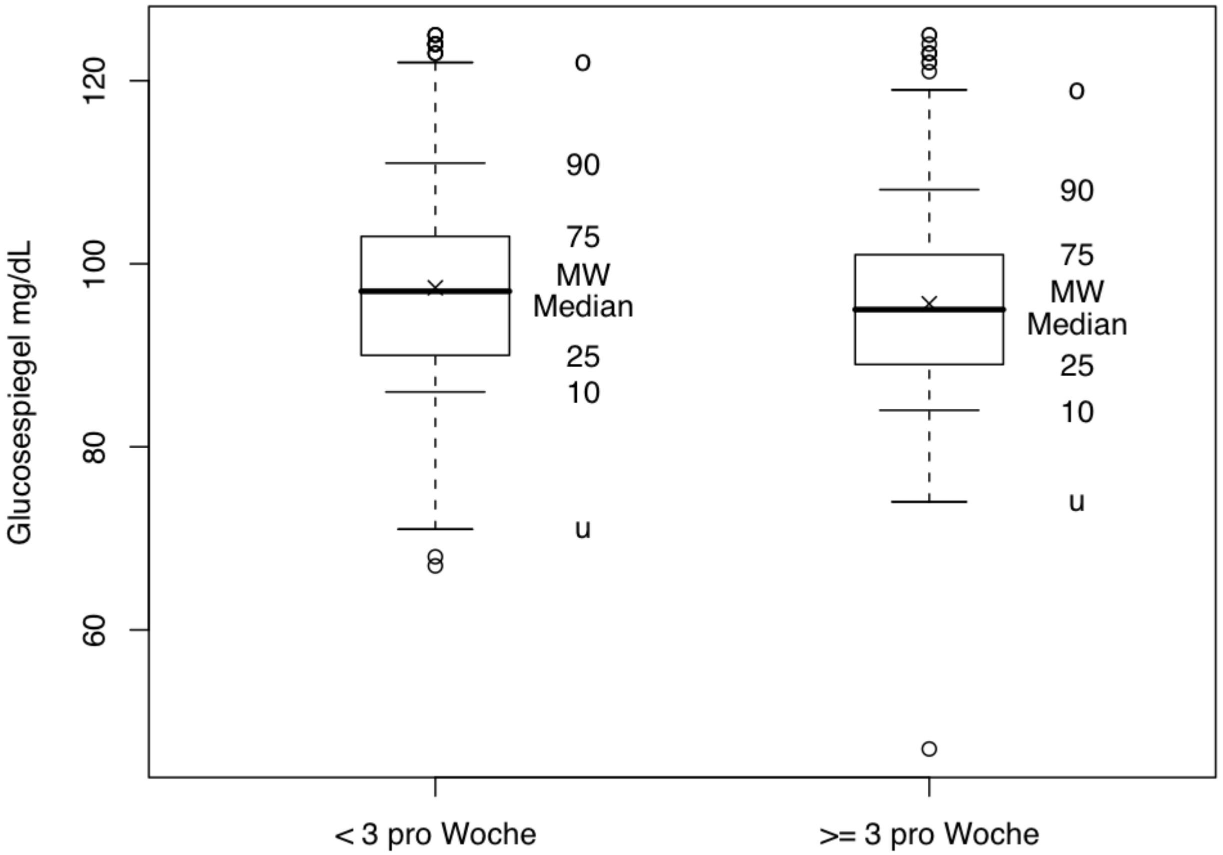


n = 1188



n = 840

Box-Plot



Plots are not standerdised

Double interquartile range

t-test (1)

- Theoretical values in the population of all women: mean blood glucose level in women without sport μ_0 and with sport μ_1
- Estimation of both parameters using the mean of the samples: $\hat{\mu}_0 = 97{,}4$ and $\hat{\mu}_1 = 95{,}6$
- In order to evaluate if the difference in the mean suggests a diffence in the population, the standerdised difference (with the estimated standard error of the difference in the means $\hat{S}E(\hat{\mu}_1 - \hat{\mu}_0)$) is used
- The result is the $t_{(n-1)}$-distributed statistic: $T = \dfrac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{S}E(\hat{\mu}_1 - \hat{\mu}_0)}$

t-test (2)

- t-Test for the effect of sport:

$$T = (95{,}6 - 97{,}4)/0.437 = -4.023;\ p < 0{,}0005$$

- t-distribution for n > 50 is approximately identical to the normal distribution

Representation as regression model (1)

- Blood glucose level = typical blood glucose level + effects of other factors (incl. physical activity) + effect of random effects
- Physical activity as covariate with values 0 (for ”no”) and 1 (for “yes”), we get *model M_1*:
- $y = \beta_0(\text{Inter}) + \beta_1(\text{sport}) * x_1 + \varepsilon$
- β_0 (Intercept): blood glucose level with “physical activity no“
- β_1 : blood glucose level **difference** with “physical activity yes” or additional affect due to sport, here (in the special case of a binary covariate) difference in groups
- ε : random error, $N(0,\sigma^2)$

Representation as regression model (2)

- Parameters β_0 and β_1 estimated from the data
- Estimates for β_1 correspond to t-test
- Estimation for model M_1:

```
         Estimate   Std. Error   t value     Pr(>|t|)
(Inter)   97.3796     0.2814     345.999      < 2e-16 ***
(Sport)   -1.7439     0.4373     -3.988       6.9e-05 ***
```

Significance vs. relevance

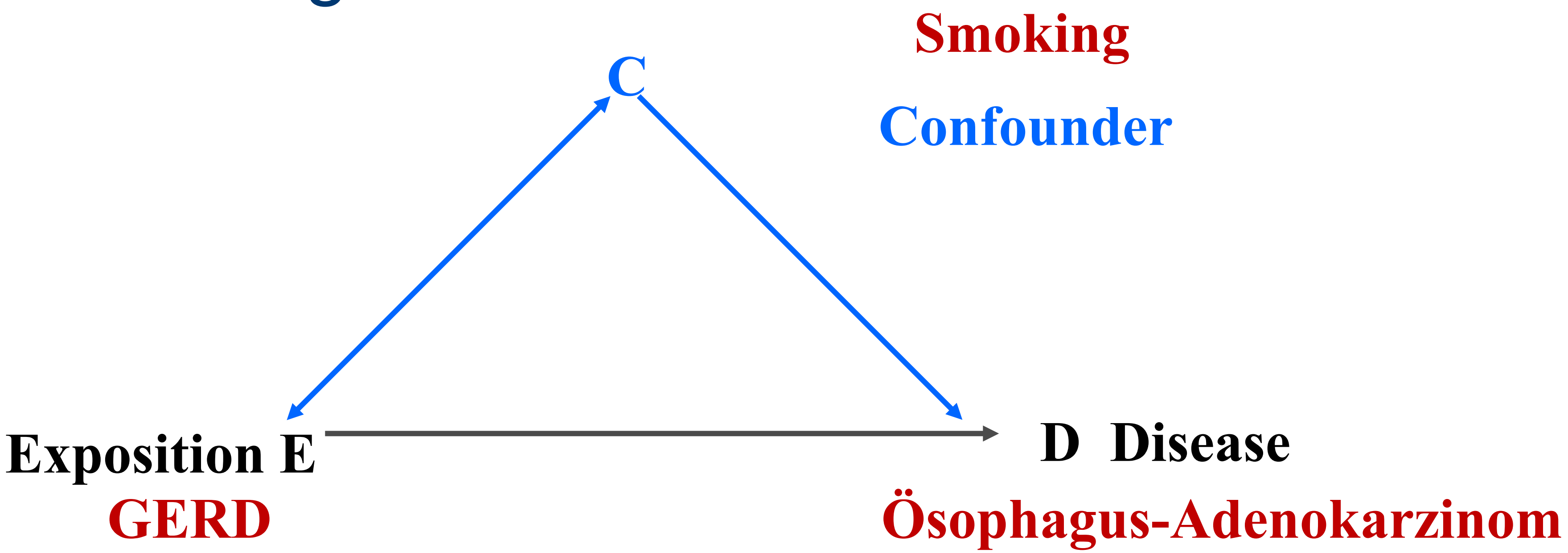
Influence of sport on blood glucose level

Statistically significant!

But clinically relevant?

Are further factors relevant?

Confounding



Smoking

Confounder

C

Exposition E

GERD

D Disease

Ösophagus-Adenokarzinom

A Confounder C is a risk factor which has an influence on the development of disease D and which is associated with exposition E. Several confounders may be present!

Lagergren et al NEJM 1999

Confounding: Fiktives Beispiel (1)

		D^+	D^-	Gesamt
C^+	E^+	30	270	300
	E^-	6	144	150
	Gesamt	36	414	450
C^-	E^+	30	30	60
	E^-	130	260	390
	Gesamt	160	290	450
Gesamt	E^+	60	300	360
	E^-	136	404	540
	Gesamt	196	704	900

$$\hat{OR} = \frac{30 \cdot 144}{6 \cdot 270} = 2{,}67$$ **> 1**

95%-Konfidenzintervall: 1,09 bis 6,57

$$\hat{OR} = \frac{30 \cdot 260}{30 \cdot 130} = 2{,}0$$ **> 1**

95%-Konfidenzintervall: 1,07 bis 3,74

$$\hat{OR} = \frac{60 \cdot 404}{300 \cdot 136} = 0{,}59$$ **< 1**

95%-Konfidenzintervall: 0,41 bis 0,86

"Simpson's Paradoxon"

HERS example - Consideration of further factors (1)

- Women who are more physically active might probably be younger, have a different level of alcohol consumption and possible a lower body-mass-index.

- The model is adjusted for these possible confounders by including them into the regression model:

blood glucose level =

$$\beta_0 + \beta_1(\text{sport}) * \text{„sport 0/1“} + \beta_2(\text{age}) * \text{age} + \beta_3(\text{alcohol}) * \text{„alcohol consumption 0/1“} + \beta_4(\text{BMI}) * \text{BMI}$$

Consideration of further factors (2)

Model M_2: $y = \beta_0 + \beta_1(\text{sport}) * x_1 + \beta_2(\text{age}) * x_2 + \beta_3(\text{alcohol}) * x_3 + \beta_4(\text{BMI}) * x_4 + \varepsilon$

- Factors age, alcohol and BMI are further confounder
- The estimated effect of „physical activity“ has to be interpreted when all other factors are fixed

Consideration of further factors (3)

- Estimation for model M_2:

	Estimate	Std. Error	t value	Pr(>|t|)
$\hat{\beta}_0$	78.96239	2.59284	30.454	<2e-16
$\hat{\beta}_1$ (Sport)	-0.95044	0.42873	-2.217	0.0267
$\hat{\beta}_2$ (Age)	0.06355	0.03139	2.024	0.0431
$\hat{\beta}_3$ (Alcohol)	0.68026	0.42196	1.612	0.1071
$\hat{\beta}_4$ (BMI)	0.48924	0.04155	11.774	<2e-16

- As a comparison: estimation without further factors (model M_1):

	Estimate	Std. Error	t value	Pr(>|t|)	
$\hat{\beta}_0$	97.3796	0.2814	345.999	< 2e-16	***
$\hat{\beta}_1$ (Sport)	-1.7439	0.4373	-3.988	6.9e-05	***

- Sport in M_2 is (absolutely) smaller!

Consideration of further factors (4)

- Both models provide an estimate (index) for the blood glucose level:

From M_1: $\hat{y}_1$ = 97.38 - 1.74 * sport

From M_2: $\hat{y}_2$ = 78.96 - 0.95 * sport + 0.06 * age +

0.68 *alcohol + 0.49 * BMI

Consideration of further factors (5)

Selection of relevant variables

- P-value of β_3 (Alcohol) 0.107, non significant for α=5%
- Eliminate alcohol from the model?
- Key issue of many analyses in all types of regression models
- Which variables to include in a model?

Several factors: adjustment of an effect

Integration of prognostic factors in regression models depends on the nature of the observed outcome

- Multiple linear regression model
- Logistic regression model
- Cox-Regression

Enables simultaneous analysis of the effect of treatment and prognostic factors

Cox-Regression

Extension of the logistic regression model for survival analysis

Consider intensity of the occurrence of an event (hazard rate) at time t:

$$\lambda(t) = \lim_{h \to 0} \frac{1}{h} P\left(t < T \leq t + h \mid T > t\right)$$

Assumption :

For two groups of patients A and B the hazard rates are proportional to each other

$$\frac{\lambda_B(t)}{\lambda_A(t)} = \text{constant } = HR \text{ (Hazard Ratio)}$$

Cox-Regression

Example: 2 explanatory variables

X_1: Prognostic factor or treatment (A or B)

X_2: Prognostic factor ("good" or "bad")

Cox regression model

$$\lambda(t) = \lambda_0(t) \cdot \exp(\beta_1 X_1 + \beta_2 X_2)$$

Interpretation of the parameters as relative risks

$$\exp(\beta_1) = HR(B : A)$$
$$\exp(\beta_2) = HR(F_2 : F_1)$$

Consider an additional interaction

$$\text{Factor } X_1 \times \text{Factor } X_2$$

$$\lambda(t) = \lambda_0(t) \cdot \exp\left(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2\right)$$

- With a test for $\beta_3 = 0$, the question of an interaction between the two factors can be investigated