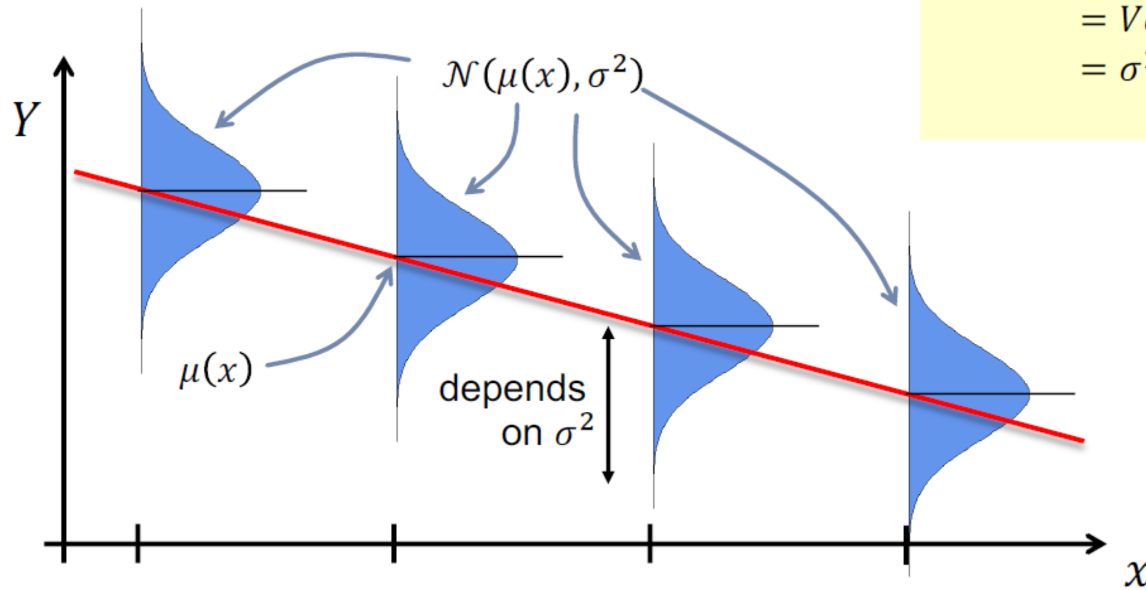
Biostatistics , Week 10

- Logistic regression for binary outcome
  - Interpretation of the coefficients
  - CI of the coefficients
  - Using logistic regression to predict (Y=1|x)

Recap: Linear regression for continuous outcomes

1. $(Y|X = x) \sim N(\underbrace{\beta_0 + \beta_1 x}_{\mu(x)}, \sigma^2)$
2. $Y = \beta_0 + \beta_1 x + \varepsilon$
   - $\varepsilon \sim N(0, \sigma^2)$

$$
\begin{aligned}
E(Y) &= E(\beta_0 + \beta_1 x + \varepsilon) \\
&= \beta_0 + \beta_1 x + E(\varepsilon) \\
&= \beta_0 + \beta_1 x \\
Var(Y) &= Var(\beta_0 + \beta_1 x + \varepsilon) \\
&= Var(\varepsilon) \\
&= \sigma^2
\end{aligned}
$$



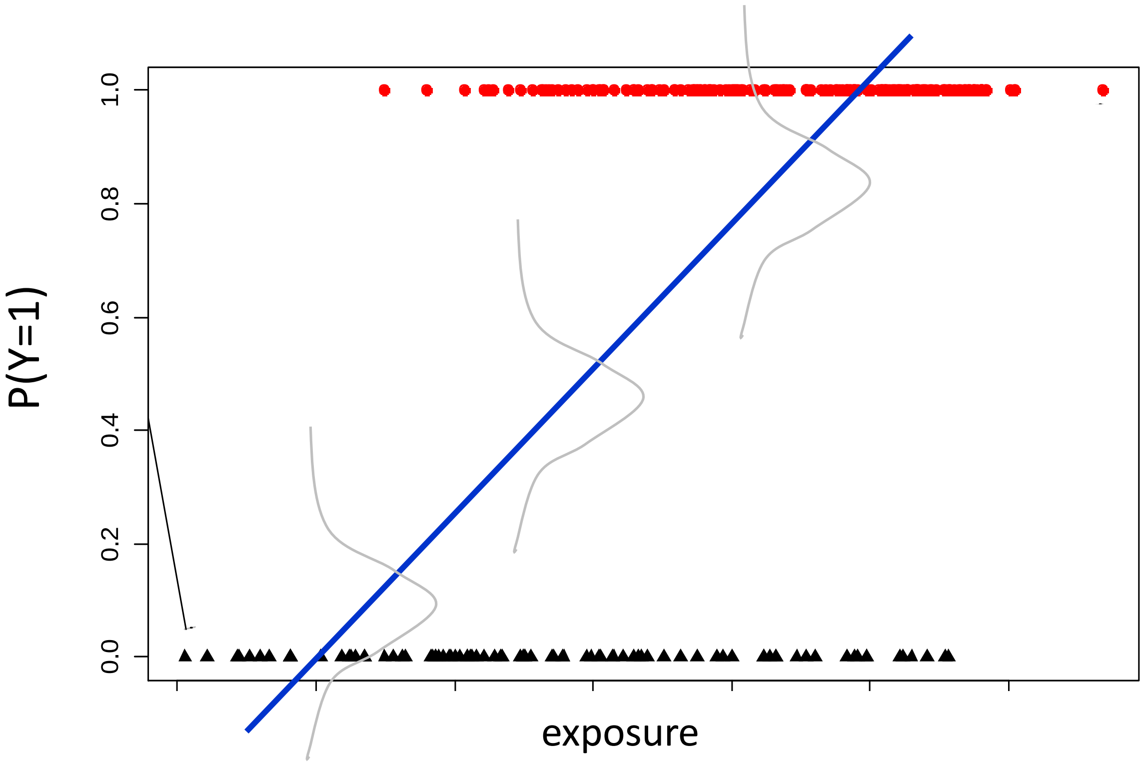
What if outcome is binary?

Logistic Regression

Why switching from linear to logistic regression?

Visualize the fitted model together with data.

```
fit = lm( y ~ exposure, data=my.dat)
```



Problems:

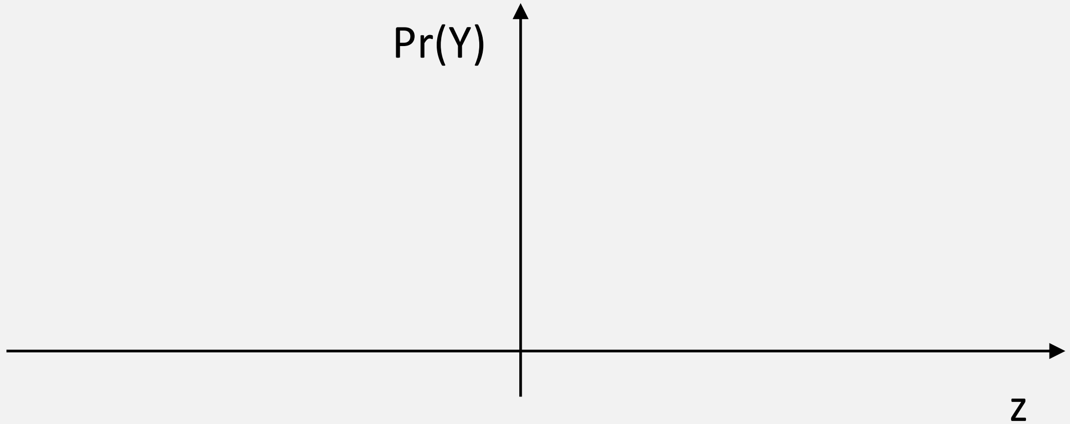
1) linear model can **yield impossible Expected values for p outside [0,1]**
2) model assumptions are violated, since residuals are <u>not</u> $\sim N(0,\sigma^2)$

Find a suitable squeezing function

- Idea of logistic regression

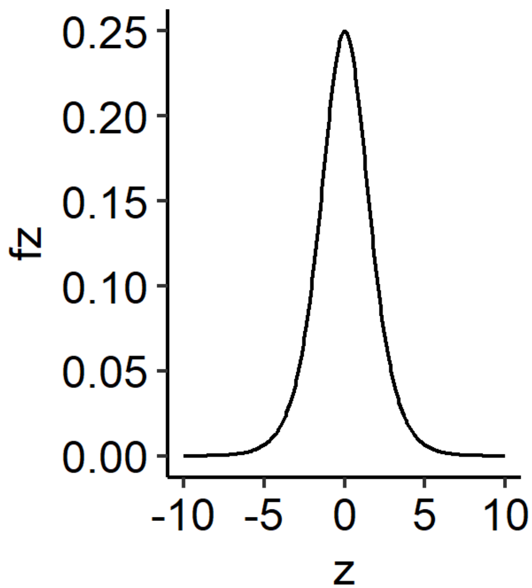
- Take output of linear regression
  - $z = a \cdot x + b \qquad z \in [-\infty, \infty]$
- and squeeze it to [0 and 1]

- **Task: Draw a function which could do that.**
  - Discuss with your neighbor



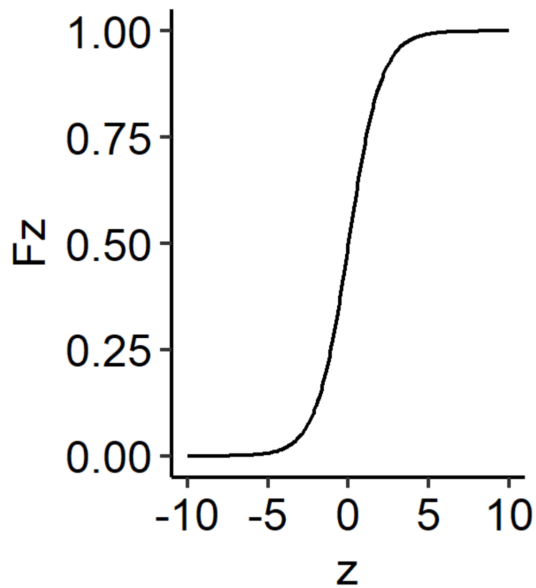
Logistic Distribution

PDF (f_Z)



$$f_Z(z) = \frac{e^z}{(1+e^z)^2}$$

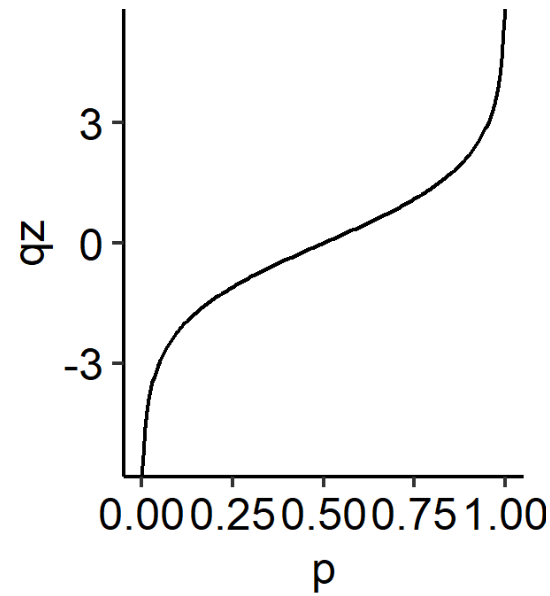
CDF (F_Z)



$$F_Z(z) = \frac{1}{1+e^{-z}}$$

$F_Z(z)$ =expit(z)
= sigmoid(z)

Quantiles function (F_Z^{-1})



$$F_Z^{-1}(p) = \log \frac{p}{1-p}$$

$$F_Z^{-1}(p) = \log(\text{odds}) = \text{logit}(p)$$

Logistic regression

Logit transformation: $$h(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \log(\text{odds})$$

Approach: probabilities are transformed to ***logged odds*** *which can then be modeled linearly.*

Logistic regression:

$$h(\pi) = \log\left(\frac{P(Y=1|\vec{X})}{1-P(\mathrm{Y}=1|\vec{X})}\right) = \underbrace{\beta_0 + \beta_1 x_1 + \ldots + \beta x_p}_{\text{Linear predictor } \eta_i}$$

link-function

Remark: the logistic regression model does not contain an error term, since the data variability is captured by the conditional Bernoulli distribution (lin reg is an exception, since only μ but not σ is modeled.

Logistic Regression

Idea:
a continuous latent (unobserved) variable determines the probability to observe $Y = 1$

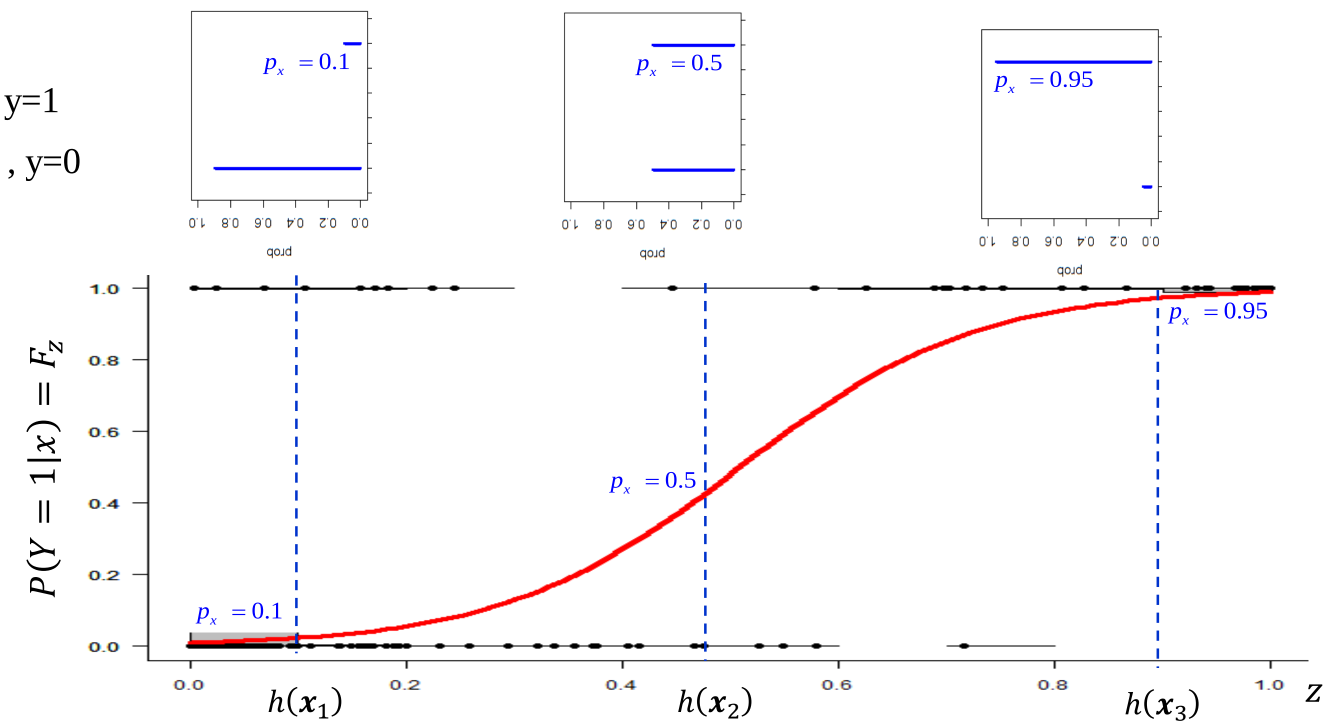
CPD: $Y_{X_i} = (Y|X_i) \sim \mathrm{Ber}(p_{x_i})$

$Y_x \in \{0,1\}$, $p_x = P(Y=1|x) \in [0,1]$

We don't model the p_x directly, but a value $h(x)$ indicating a point in the latent variable z yielding via the CDF F_Z the p_x:

$$p_x = F_Z(h(\boldsymbol{x}))$$

$$P(Y|X=x) = \begin{cases} p_x & , y=1 \\ 1-p_x & , y=0 \end{cases}$$



The conditional expected value in a logistic regression model

Logistic Regression Model

- The binary (Y|x) has a conditional Bernoulli distribution B(π(x)).
- The parameter of this distribution is π(x), the disease probability which might be different for each subject depending on its co-variables X.

Now please note that: $\pi_i(x) = P(Y_i = 1 | x) = E[Y_i | x]$

→ A common notion of the logistic regression model is to see it as a model where we try to find a relation between the probability for the outcome "1" (= the expected value of the binary response Y) and the predictors $X_1,\ldots X_p$

Important: linear regression is not appropriate!

Estimating the coefficients in logistic regression

$$\log\left(\frac{P(Y=1|X)}{1-P(\mathrm{Y}=1|X)}\right)=\beta_0+\beta_1 x_1+\ldots+\beta x_p \qquad P(Y=1|X)=\frac{\mathrm{e}^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+\mathrm{e}^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

$$P(\mathrm{Y}_i=1|X_i)=\pi_i\ ,\ P(\mathrm{Y}_i=0|X_i)=1-\pi_i \xrightarrow[\text{coding } Y\in\{0,1\}]{\text{useful notation}} P(\mathrm{Y}_i|X_i)=\pi_i^{\,Y_i}\cdot(1-\pi_i)^{1-Y_i}$$

We estimate the coefficients by maximizing the Likelihood

(the coefficients β are contained in π since π is determined as a function of the linear predictor)

$$L(\beta)=\prod_{i=1}^{n}\mathrm{P}(\mathrm{Y}_i|\mathrm{X}_i) \qquad \mathrm{l}(\beta)=\sum_{i=1}^{n} y_i\log(\pi_i)+(1-y_i)\log(1-\pi_i)$$

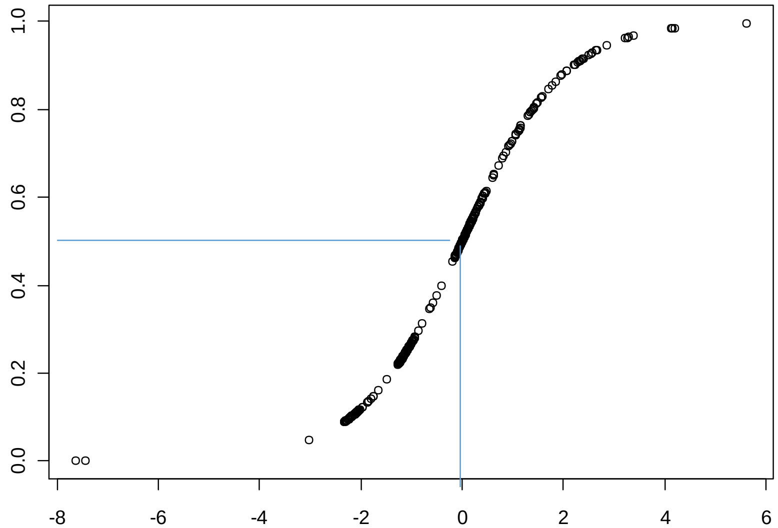
There is no closed formula for the MLE of the coefficient since they are determined in an iterative approach (IRLS).

Back transformation to probabilities

log-odds: $\log(\text{odds}) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$

Probability: $P(Y = 1|x) = \dfrac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p)}}$

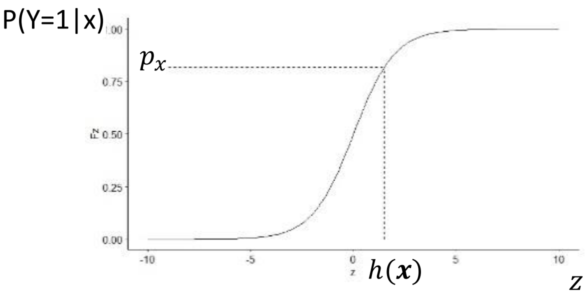
$$P(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p)}}$$



$$\log(\text{odds}) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

Interpretation of the coefficients in a logistic regression model

$$\log(\text{odds}(\boldsymbol{p}(\boldsymbol{x}))) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$



The **coefficient β_k** as the **log-odds-ratio** for $Y = 1$ when comparing a situation where x_k is increases by 1 unit (while fixing all other variables) with the situation before increasing x_k

$$\log(\text{OR}_k) = \log\left(\frac{\text{odds}(x_1,\ldots,x_k+1,\ldots,x_p)}{\text{odds}(x_1,\ldots,x_k,\ldots,x_p}\right) = \log\left(\frac{e^{\beta_0}\cdot e^{\beta_1 x_1}\cdot\ldots\cdot e^{\beta_k(x_k+1)}\ldots\cdot e^{\beta_p x_p}}{e^{\beta_0}\cdot e^{\beta_1 x_1}\cdot\ldots\cdot e^{\beta_k x_k}\ldots\cdot e^{\beta_p x_p}}\right) = \log\left(e^{\beta_k}\right) = \beta_k$$

$$\Rightarrow \quad e^{\beta_k} = \text{OR}_{x_k \to x_k+1}$$

Interpretation of the coefficients in logistic regression

Interpretation of Regression Coefficient (β_k): $\qquad \overset{\wedge}{\text{response}}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{i1} + \ldots + \hat{\beta}_p \cdot x_{ip}$

- In logistic regression, we have the following relationship:

$$\log\left(\overset{\wedge}{odds}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p$$

- In linear regression, the slope coefficient β_k gives the difference in the expected response (here: `log(odds)`) as x_k increases by 1 unit (while fixing all other variables)

$$\beta_k = \left(\log(\text{odd}_{x_k+1}) - \log(\text{odd}_{x_k})\right) = \log\left(\frac{\text{odd}_{x_k+1}}{\text{odd}_{x_k}}\right) = \log(\text{OR}_{x_k \to x_k+1})$$

$$\Rightarrow \quad e^{\beta_k} = \text{OR}_{x_k \to x_k+1}$$

- Thus e^{β_k} gives the OR when the risk factor x_k is increased by 1 unit (and all other variables hold fix)

- If β_k=0, the odds (and probability) is equal at all x_k levels (e^{β_k} =1)
- If β_k>0 , the odds (and probability) increases as x_k increases (e^{β_k} >1)
- If β_k< 0 , the odds (and probability) decreases as x_k increases (e^{β_k} <1)

Confidence intervals for the coefficients in logistic regression

- Note, that the coefficients β in logistic regression
  - are, asymptotically, normally distributed
  - Don't have a t distribution (as in linear regression)
- Therefore quantiles of the N(0,1) distribution are used
- For a 95% confidence interval for β

$$\hat{\beta} \pm 1.96 \cdot \mathrm{se}(\hat{\beta})$$

Confidence intervals for the odds ratio

Determine the 95% CI for $\beta_k = \left(\log(\text{odd}_{x_k+1}) - \log(\text{odd}_{x_k})\right) = \log\left(\frac{\text{odd}_{x_k+1}}{\text{odd}_{x_k}}\right) = \log(\text{OR}_{x_k \to x_k+1})$

$$95\%\ \text{CI for } \log(\text{OR}_{x_k \to x_k+1}):\ \left[\hat{\beta}_k - 1.96 \cdot se(\hat{\beta}_k)\ ,\ \hat{\beta}_k + 1.96 \cdot se(\hat{\beta}_k)\right]$$

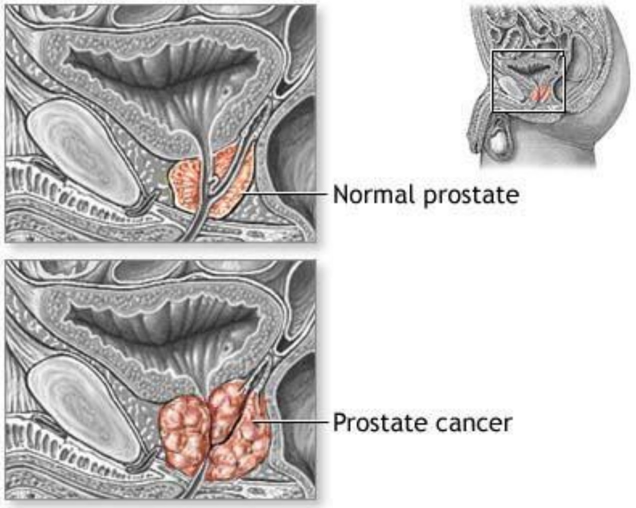
$$95\%\ \text{CI for } \text{OR}_{x_k \to x_k+1}:\ \left[e^{\hat{\beta}_k - 1.96 \cdot se(\hat{\beta}_k)}\ ,\ e^{\hat{\beta}_k + 1.96 \cdot se(\hat{\beta}_k)}\right]$$

A confidence interval for β = logOR is symmetric

A confidence interval for $OR = e^{\beta}$ is skewed - the further OR is from 1, the more skewed is the confidence interval.

Example: Prostate Cancer

From a biopsy medical doctors can determine different continuous scores which might help to predict if the cancer is aggressive or not.



Target variable Y is binary (0 or 1):

Y_i = 1 if tumor is aggressive
Y_i = 0 if is not aggressive

Predictor or Covariates variables X_i are continuous

X_1: PSA: concentration of a predictive antigen
X_2: GS: gleason score is grading the cells

Goal:
We'd like to formulate a model that predicts the Y-value from the values of the two predictors or a probability for Y=1.

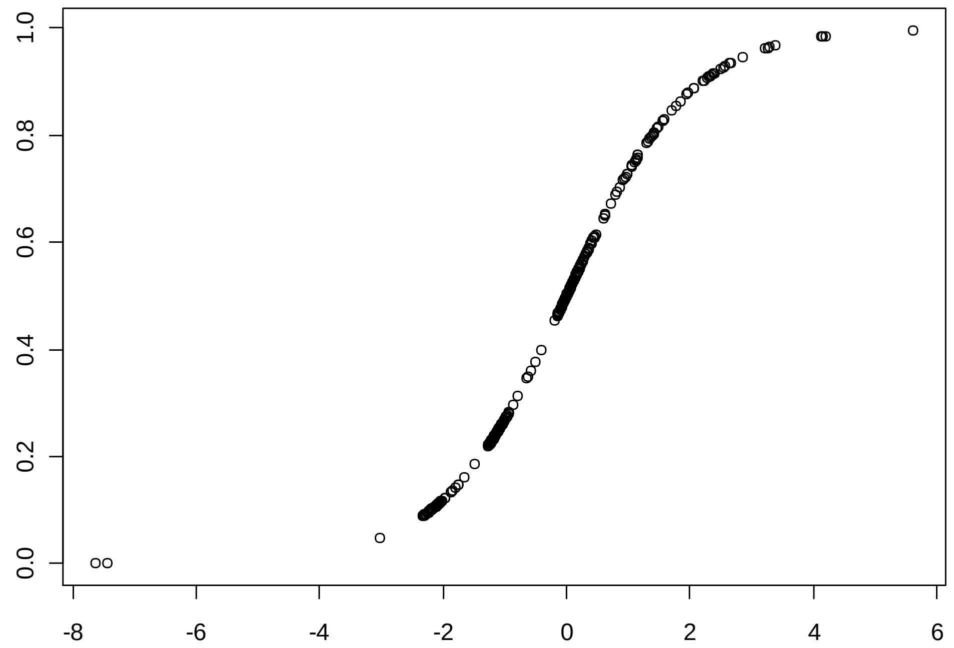
Modeling the prostate data by logistic regression in R

```
> pros1.reg <- glm(y ~ psa + gs, family=binomial)
> summary(pros1.reg)
Call:
glm(formula = y ~ psa + gs, family = binomial)
Deviance Residuals:
    Min        1Q    Median        3Q       Max
-2.2100   -0.7692   -0.4723   1.0431   2.1398

Coefficients:
             Estimate   Std. Error  z value   Pr(>|z|)
(Intercept) -7.639296   1.011128    -7.555   4.18e-14 ***
psa          0.026677   0.008929     2.988   0.00281 **
gs          1.059344   0.158327     6.691   2.22e-11 ***
```

$$\log\left(\hat{odds}\right) = \log\left(\frac{\hat{\pi}}{1-\pi}\right) = -7.6 + 0.027 \cdot \text{psa} + 1.06 \cdot \text{gs}$$

Logistic regression model for the prostate data

$$P(Y=1|x) = \frac{1}{1+e^{-(-7.6+0.027\cdot \text{psa}+1.06\cdot \text{gs}}}$$



$$h = \text{lin.predictor} = -7.6 + 0.027 \cdot \text{psa} + 1.06 \cdot \text{gs}$$

Interpretation of the R output in the prostate example

$$\log\left(\hat{odds}\right) = \log\left(\frac{\hat{\pi}}{1-\pi}\right) = -7.6 + 0.027 \cdot \text{psa} + 1.06 \cdot \text{gs}$$

Which meaning has the estimated `gs`-coefficient 1.06?

- for two men who have same psa but a difference in their `gs` of one unit the expected `difference of log(odds)` or `logOR` is 1.06.

We usually look at the exponential of the coefficient:

- $\exp(\beta_2)$ = exp(1.06) = 2.88
- The expected odds for having an aggressive tumor is `2.88`-times higher for man with `gs=7` than for a man with `gs=6`
- The odds ratio OR for a 1 unit difference in **`gs`** is

  $OR_{gs\text{->}gs+1} = e^{1.06} = 2.88$

- You also need to interpret changes as 'adjusting for PSA'

Interpretation of the coefficients in logistic regression
Case 1: Continuous explanatory variable (reminder)

Let X_k be a continuous explanatory variable (e.g. PSA-value in the prostate example)

Interpretation of Regression Coefficient : $\log\left(\widehat{odds}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p$

– In logistic regression, the intercept (β_0) gives the log-odds of *Y=1 (disease)* amongst the subjects who have the value zero in all explanatory variables (*)

– the slope coefficient β_k gives the difference in the expected response (here: `log(odds)`) as x_k increases by 1 unit (while fixing all other variables)

$$\beta_k = \left(\log(\mathrm{odd}_{x_k+1}) - \log(\mathrm{odd}_{x_k})\right) = \log\left(\frac{\mathrm{odd}_{x_k+1}}{\mathrm{odd}_{x_k}}\right) = \log(\mathrm{OR}_{x_k \to x_k+1})$$

$$\Rightarrow \quad e^{\beta_k} = \mathrm{OR}_{x_k \to x_k+1}$$

– Thus e^{β_k} gives the OR when the risk factor x_k is increased by 1 unit (and all other variables hold fix)

(*) this interpretation is valid for cohort or cross-sectional studies, but not for case-control studies where #D is fixed and therefore the odds for disease cannot be estimated from the collected data

Interpretation of the coefficients in logistic regression
Case 2: Binary explanatory variable

Let X_1 be a binary explanatory variable. $\log\left(\widehat{odds}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1$

We first consider one binary exposure variable X_1 taking on only two values, say, x_1=1 (exposed) and X_1=0 (unexposed).

First consider X_1=0: $$\log\left(\widehat{odds}(X_1 = 0)\right) = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 = \hat{\beta}_0$$

$\Rightarrow$ The intercept (β_0) gives the log odds of *Y=1 (disease)* amongst the subjects in the reference level X_1=0 meaning here among the unexposed (*).

Now consider X_1=1: $$\log\left(\widehat{odds}(X_1 = 1)\right) = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 = \hat{\beta}_0 + \hat{\beta}_1$$

$$\Rightarrow \log\left(\mathrm{OR}_{x_1=0 \to x_1=1}\right) = \log\left(\mathrm{odds}(X_1 = 1)\right) - \log\left(\mathrm{odds}(X_1 = 0)\right)$$
$$= \hat{\beta}_0 + \hat{\beta}_1 - \hat{\beta}_0 = \hat{\beta}_1$$

Thus e^{β_1} gives the OR when the risk factor X_1 goes from the reference level (X_1=0, unexposed) to the observed level (X_1=1, exposed). It is essential to choose a good reference level (many observations with typical value.

(*) this interpretation is valid for cohort or cross-sectional studies, but not for case-control studies

Interpretation of the coefficients in logistic regression
Case 1: Categorical explanatory variable (>2 levels)

Consider a categorical explanatory variable with k>2 different levels.
(e.g. `drug` with 4 levels: `alcohol, heroin, cannabis, ecstasy`)

For an exposure with `K` distinct levels, one level is first chosen as the baseline or reference group (e.g. alcohol). Refer to that level as level 0.
Then the following `K-1` indicator variables X_i are defined, which allow to specify the exposure level of a subject:

X_1=1 if an individual's exposure is at level 1, and X_1=0 otherwise.
X_2=1 if an individual's exposure is at level 2, and X_2=0 otherwise.
…
X_{K-1}=1 if an individual's exposure is at level $K-1$, and X_{K-1}=0 otherwise.

$$\log\left(\hat{odds}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_{k-1} X_{k-1}$$

e^{β_j} is the OR comparing exposure level j to the reference level 0.

$e^{\beta_4 - \beta_2}$ is the OR comparing exposure level 4 to exposure level 2.

The pancreatic cancer and coffee drinking example

We take the case-control study of coffee drinking and pancreatic cancer (MacMahon et al., 1981) as an example for the further discussion.

crude table

		Pancreatic Cancer		
		Cases	Controls	
Coffee drinking (cups per day)	>1	347	555	902
	0	20	88	108
		367	643	1010

stratified table

		Coffee Drinking (Cups per Day)				
Sex	Disease Status	0	1–2	3–4	≥5	Total
Men	Case	9	94	53	60	216
	Controls	32	119	74	82	307
Women	Case	11	59	53	28	151
	Controls	56	152	80	48	336
	Total	108	424	260	218	1010

Ungrouped and grouped data

When collecting data, the observed variable values corresponding to one study subject (or patient) are typically collected in one row of the original data set. This is called ***ungrouped*** data.

If the study subjects have many identical values in their variables, which is often the case if the variables are categorical, a more compact way of representing the data by forming groups with identical variable setting. This is as "***grouped data***".

Coffee data in ungrouped and grouped representation

Ungrouped data – 1010 x 4
1 row for each patient

	case	coffee	sex	index.coffee
1	0	3	0	1
2	0	3	0	1
3	0	[illegible]	[illegible]	[illegible]
4	0	[illegible]	[illegible]	[illegible]
5	0	3	[illegible]	[illegible]
6	0	3	0	1
7	0	3	0	1
8	0	3	0	1

⋮

81	0	3	[illegible]	[illegible]
82	0	3	[illegible]	[illegible]
83	0	2	0	1
84	0	2	0	1
85	0	2	0	1
86	0	2	0	1
87	0	2	0	1

⋮

Case:
0=ctrl (no cancer)
1=case (cancer)

Coffee:
Level of coffee drinking

Sex:
0=male
1=female

Grouped data: – 8 x 4
1 row for each risk factor setting defining a patient-group

(8 groups = 2 Gender x 4 coffee-levels)

	Coffee	Sex	Count.Ctrl	Count.Case	Count
1	0	0	32	9	41
2	0	1	56	11	67
3	1	0	119	94	213
4	1	1	152	59	211
5	2	0	74	53	127
6	2	1	80	53	133
7	3	0	82	60	142
8	3	1	48	28	76

Count.Ctrl:
0-outcome in this group

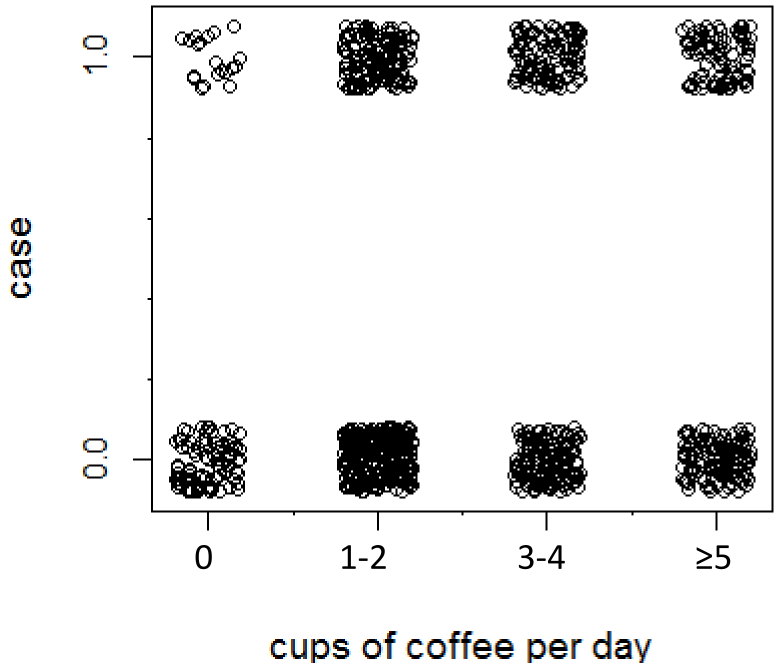
Count.Case:
1-outcome in this group

Count n_i in ith group:
#patients with this variable setting

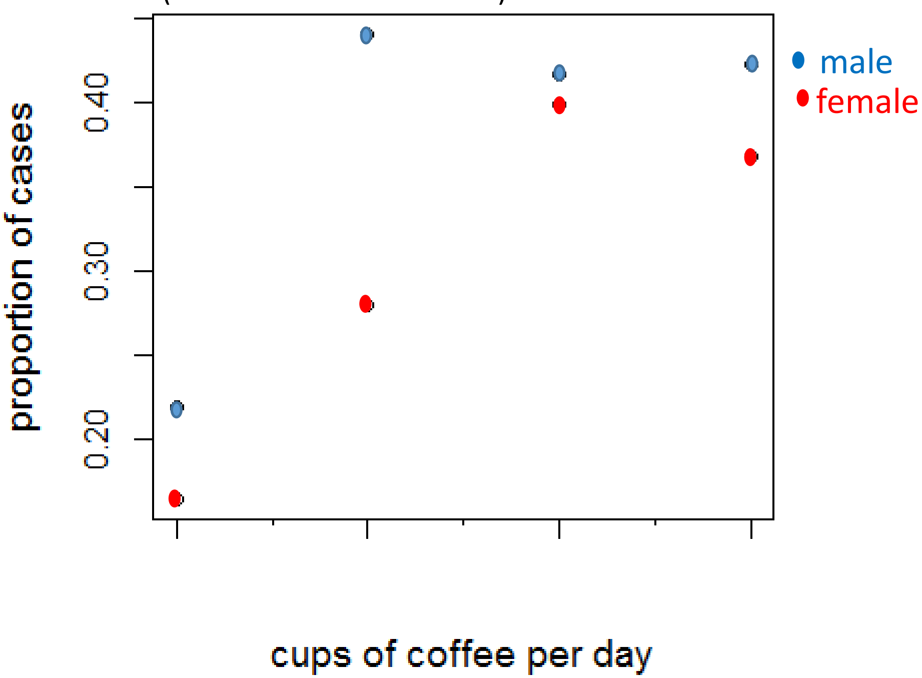
How to get from grouped to ungrouped data using R
Coffee data in ungrouped and grouped representation

```
co.gr <- read.csv("coffee.csv", sep=";")

co.ungr = data.frame(case = as.factor(unlist(mapply(rep, x=co.gr$Case, times=co.gr$Count))))
co.ungr$coffee = unlist(mapply(rep, x=co.gr$Coffee, times=co.gr$Count))
co.ungr$sex = as.factor(unlist(mapply(rep, x=co.gr$Sex, times=co.gr$Count)))
co.ungr$index.coffee=as.factor(1*(co.ungr$coffee>0))
```

Ungrouped data contain 1010 rows – one per patient (outcome 0/1)



Grouped data contain 8 rows – one per patient group (2Gender x 4coffeelevels)



Logistic regression using the ungrouped data

```
> fit.ungr=glm(case~as.factor(coffee), family=binomial(link="logit"),data=co.ungr)
> summary(fit.ungr)

Call:
glm(formula = case ~ as.factor(coffee), family = binomial(link = "logit"),
    data = co.ungr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0234  -1.0168  -0.9462   1.3470   1.8365

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.4816     0.2477  -5.981 2.22e-09 ***
as.factor(coffee)1   0.9099     0.2676   3.401 0.000672 ***
as.factor(coffee)2   1.1081     0.2780   3.986 6.73e-05 ***
as.factor(coffee)3   1.0914     0.2836   3.849 0.000119 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1323.8  on 1009  degrees of freedom
Residual deviance: 1303.6  on 1006  degrees of freedom
AIC: 1311.6

Number of Fisher Scoring iterations: 4
```

Do not over-interpret these p-values, but compare models with and w/o this covariate (via $\chi 2$ test using the R function anova) to asses if this covariate is significant.

#steps in IRLS estimation

Logistic regression using the grouped data

```
> fit.gr=glm(cbind(Count.Case,Count.Ctrl)~as.factor(Coffee),
+              family=binomial(link="logit"),data=co.gr)
> summary(fit.gr)

Call:
glm(formula = cbind(Count.Case, Count.Ctrl) ~ as.factor(Coffee),
    family = binomial(link = "logit"), data = co.gr)

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.5534  -0.4495   2.4125  -2.5053   0.2206  -0.2161   0.4571  -0.6296

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.4816     0.2477  -5.981 2.22e-09 ***
as.factor(Coffee)1   0.9099     0.2676   3.401 0.000672 ***
as.factor(Coffee)2   1.1081     0.2780   3.986 6.73e-05
as.factor(Coffee)3   1.0914     0.2836
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33.469  on 7  degrees of freedom
Residual deviance: 13.306  on 4  degrees of freedom
AIC: 61.257

Number of Fisher Scoring iterations: 4
```

We get the same estimates & se with grouped or ungrouped data

The resulting deviances are different with grouped or ungrouped data

Comparing two nested models in R

```
> anova(fit.gr,fit.gr2,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(Count.Case, Count.Ctrl) ~ as.factor(Coffee)
Model 2: cbind(Count.Case, Count.Ctrl) ~ as.factor(Coffee) + Sex
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         4     13.306
2         3      4.268  1   9.0378 0.002644 **
```

Adding of the explanatory variable `Sex` improves the model fit.

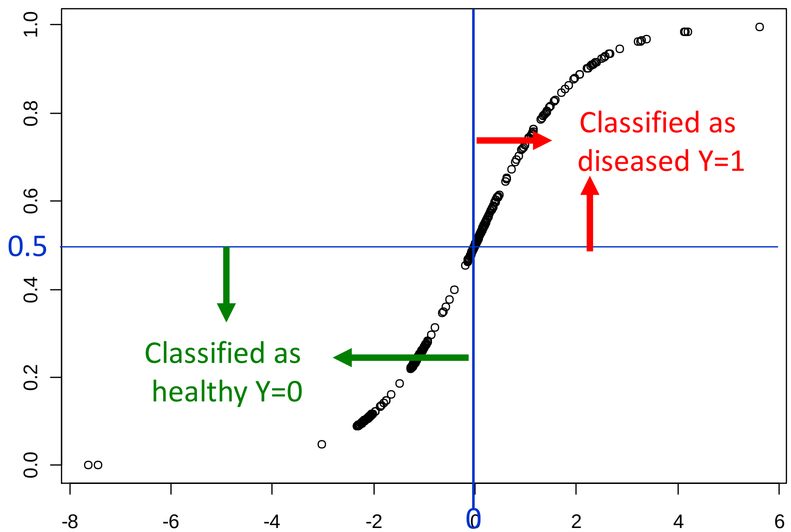
```
> drop1(fit.gr2, test="Chisq")
Single term deletions

Model:
cbind(Count.Case, Count.Ctrl) ~ as.factor(Coffee) + Sex
                  Df Deviance    AIC     LRT  Pr(>Chi)
<none>                  4.268 54.219
as.factor(Coffee)  3   21.870 65.822 17.6024 0.0005312 ***
Sex                1   13.306 61.257  9.0378 0.0026445 **
```

Using logistic regression for binary classification

A logistic regression model is not only used explaining the relation between predictors and outcome which is captured in the coefficients –
we can also use it to make predictions based on the predictor values.

Fitted probability: $\hat{\pi}_{\mathrm{i}} = \mathrm{g}^{-1}(\hat{\eta}_i) = \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}} = \frac{\mathrm{e}^{\hat{\beta}_0+\hat{\beta}_1 X_1+\ldots+\hat{\beta}_p X_p}}{1+\mathrm{e}^{\hat{\beta}_0+\hat{\beta}_1 X_1+\ldots+\hat{\beta}_p X_p}}$

$$\hat{\pi} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}}$$



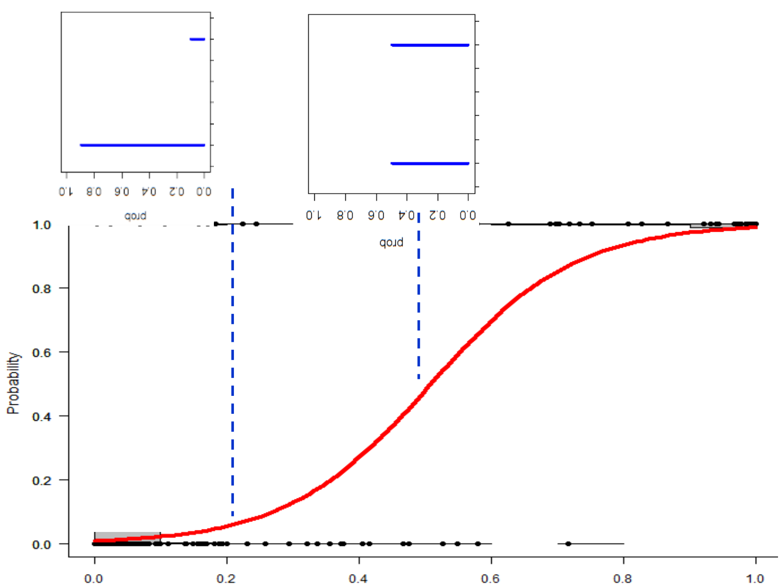
$$\hat{\eta} = \log(\widehat{\text{odds}}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p$$

Comparison Logistic Regression / Linear

Logistic Regression

$$p(x) = a \cdot x + b$$
$$Y \sim Bern(p(x))$$

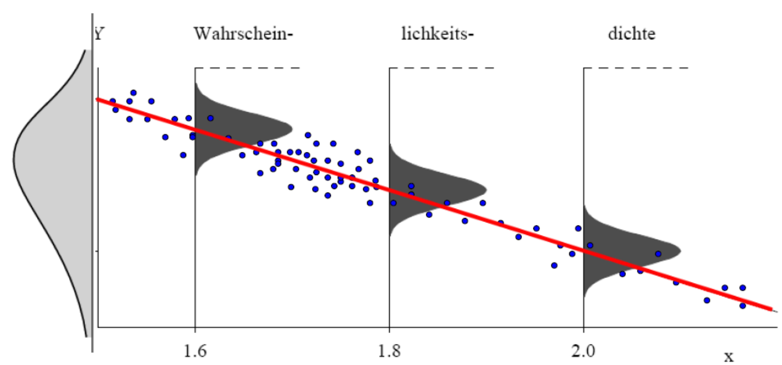
```
glm(y ~ ., binomial(logit))
```



Linear Regression

$$\mu(x) = a \cdot x + b$$
$$Y \sim N(\mu(x), \sigma = 1)$$

```
glm(y ~ ., gaussian(identity))
```



Summary

- Logistic regression can model the association between a binary outcome and a exposure and allows to adjust for confounders of any data type
  - The coefficients in a logistic regression model can be interpreted as log-OR when the corresponding predictor increases by one and all other predictors stay constant
  - Bernoulli logistic regression works on ungrouped data (1 observation = 1 row in data matrix)
  - Binomial logistic regression works on grouped data (only possible with categorical predictors)
  - When working with factor variables it is important to choose a good reference level