Generalized Linear Model

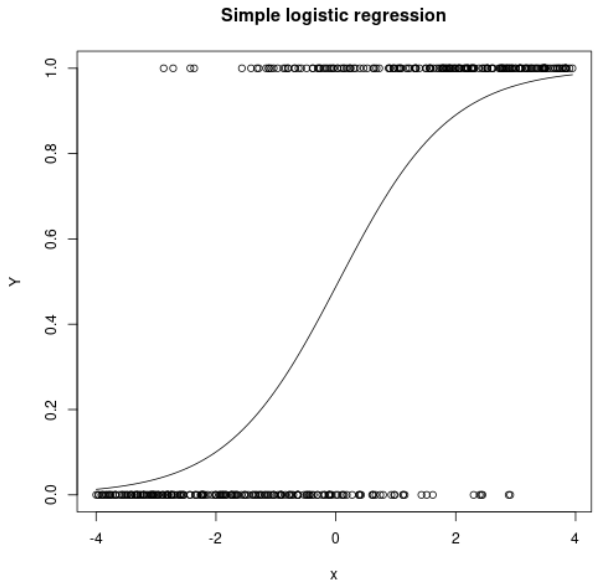
Kolloquium für Statistik

Departement of Health Professions
Bern University of Applied Sciences

October 2, 2024

Generalized Linear Model (GLM)

- We want to generalize the linear model to discrete or continuous outcomes
- Dichotomous event outcome, leading to Logistic regression
- Counts as outcome, leading to Poisson regression



Aspects of generalization

1. Link function
2. Variance function
3. Other distributions

Link function

The most important aspect is the link-function.

- Systematic part: The expectation of the response,

$$\mu_i = \mathrm{E}(Y_i),$$

  is transformed with a link function.
- The transformed expectation is called the linear predictor $\eta_i = \boldsymbol{x}_i^T \beta$.
- with the link function $h(\cdot)$, we have

$$\boxed{h(\mu_i) = \eta_i = \boldsymbol{x}_i^T \beta.} \tag{1}$$

Important link functions

- Linear regression: **Identity** function: $h(\mu_i) = \mu_i$
- Logistic regression: **logit** function: $h(\mu_i) = \text{logit}\, \mu_i$
- Poisson regression: **log** function: $h(\mu_i) = \log \mu_i$

Variance function

- Random part: The variance $\text{Var}(Y_i)$ is now a function of the expectation,

$$\text{Var}(Y_i) = \phi v(\mu_i), \tag{2}$$

where

  - $v(\cdot)$ is the variance function and
  - ϕ is the dispersion parameter, which has to be estimated or not.

Important variance functions

- Linear regression: $v(\mu_i) = 1$ with $\phi = \sigma^2$
- Logistic regression: $v(\mu_i) = \mu_i(1 - \mu_i)$ and $\phi = 1$
- Poisson regression: $v(\mu_i) = \mu_i$ and $\phi = 1$

Distributions

Each class of a GLM follows a model with density of the so-called exponential family. Special cases and most often used distributions of the exponential family are:

- The Normal distribution in Linear regression (What we have done so far)
- The Binomial distribution in Logistic regression
- The Poisson distribution in Poisson regression

Recap: Linear Model

- Model:

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

- The expectation μ_i is

$$\mu_i = \mathrm{E}(Y_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}. \tag{4}$$

- The link function $h(\cdot)$ is the identity and the variance function is $v(\mu_i) = 1$, the dispersion parameter is known, $\phi = \sigma^2$.
- Interpretation: β_j is the difference in expectations for two subpopulations that differ on x_j by on unit (slope).

Recap: Linear Model for Fertility

- We have seen least squares estimation `lm()`

```
m.lm <- lm(Fertility ~ ., swiss)
m.lm0 <- lm(Fertility ~ 1, swiss)  ## null model fit for later
summary(m.lm)
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = swiss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.274  -5.262   0.503   4.120  15.321
##
## Coefficients:
##                  Estimate Std. Error t value   Pr(>|t|)
## (Intercept)       66.9152    10.7060    6.25 0.00000019
## Agriculture       -0.1721     0.0703   -2.45     0.0187
## Examination       -0.2580     0.2539   -1.02     0.3155
## Education         -0.8709     0.1830   -4.76 0.00002431
## Catholic           0.1041     0.0353    2.95     0.0052
## Infant.Mortality   1.0770     0.3817    2.82     0.0073
##
## Residual standard error: 7.17 on 41 degrees of freedom
## Multiple R-squared:  0.707, Adjusted R-squared:  0.671
## F-statistic: 19.8 on 5 and 41 DF,  p-value: 5.59e-10
```

The same model as GLM

- Now estimated with maximum likelihood: `glm()`
- We have to fix the distribution, here `family=gaussian`
- Now switch between this slide and the former.

```
m.glm <- glm(Fertility ~ ., swiss, family = gaussian)
m.glm0 <- glm(Fertility ~ 1, swiss, family = gaussian)  ## null model fit for later
summary(m.glm)
```

```
##
## Call:
## glm(formula = Fertility ~ ., family = gaussian, data = swiss)
##
## Coefficients:
##                  Estimate Std. Error t value   Pr(>|t|)
## (Intercept)       66.9152    10.7060    6.25 0.00000019
## Agriculture       -0.1721     0.0703   -2.45     0.0187
## Examination       -0.2580     0.2539   -1.02     0.3155
## Education         -0.8709     0.1830   -4.76 0.00002431
## Catholic           0.1041     0.0353    2.95     0.0052
## Infant.Mortality   1.0770     0.3817    2.82     0.0073
##
## (Dispersion parameter for gaussian family taken to be 51.3)
##
##     Null deviance: 7178  on 46  degrees of freedom
## Residual deviance: 2105  on 41  degrees of freedom
## AIC: 326.1
##
## Number of Fisher Scoring iterations: 2
```

What is different between `lm()` and `glm()` output?

- "Deviance" versus Sum of Squares
- "Likelihood ratio tests" versus F-tests
- Least squares `lm()`

```
anova(m.lm0, m.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Fertility ~ 1
## Model 2: Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##   Res.Df  RSS Df Sum of Sq    F  Pr(>F)
## 1     46 7178
## 2     41 2105  5      5073 19.8 5.6e-10
```

- Maximum likelihood, `glm()`

```
anova(m.glm0, m.glm, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Fertility ~ 1
## Model 2: Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        46       7178
## 2        41       2105  5     5073   <2e-16
```

Estimation and Tests

- Estimation via Maximum Likelihood
- log-likelihood $l(\beta)$: (logarithmic) probability of the data as function of the parameter vector.
- The log-likelihood l is[1]

$$l(\beta) = \sum_{i=1}^{n} \log \Pr(Y_i = y_i \mid \boldsymbol{x}_i, \beta) \tag{5}$$

- The β that maximizes $l(\beta)$ is called the Maximum Likelihood Estimate (MLE) $\hat{\beta}$
- One can show that the MLE $\hat{\beta}$ has an asymptotic normal distribution.

[1] Remember that $\log \prod_{i=1}^{n} p_i = \sum_{i=1}^{n} \log p_i$.

Estimation and Tests

- Residual Deviance "replaces" the residual sum of squares and is defined as

$$D = 2(l_{max} - l(\hat{\beta})) \tag{6}$$

where

  - l_{max} is the log-likelihood for the "maximal", the saturated model (one parameter for each observation i (the best possible fit))
  - $l(\hat{\beta})$ is log-likelihood of the MLE.

- The factor 2 is necessary for D to have a χ^2-distribution with $n - p$ degrees of freedom.
- Null Deviance replaces the total sum of squares

$$D = 2(l_{max} - l_0) \tag{7}$$

where

  - l_0 is the log-likelihood for the null model

Estimation and Tests: Likelihood-Ratio-Test

Assume two nested models *Large* and *Small*:

- The difference in deviance

$$2(l_{Large} - l_{Small}) = 2 \log \frac{L_{Large}}{L_{Small}} \tag{8}$$

- can be shown to have an asymptotic chi-square distribution with the difference of the number of parameters as degrees of freedom,

$$\boxed{2(l_{Large} - l_{Small}) \overset{approx}{\sim} \chi^2_{p_{Large} - p_{Small}}} \tag{9}$$

- H_0: Model small with p_{Small} parameters is true.
- H_1: Model large with $p_{Large} > p_{Small}$ parameters is true.
- $2(l_{Large} - l_{Small}) \overset{approx}{\sim} \chi^2_{p_{Large} - p_{Small}}$

This is the very important Likelihood-Ratio-Test.

Logistic regression

- Important and frequent model in Health Sciences.
- We have a dichotomous response variable Y_i:
  - Yes-No
  - healthy-diseased
  - etc.
- We want to model the probability of the event.
- The distribution of the Y_i is binomial with parameters π_i and $n = 1$ (bernoulli),

$$Y_i \sim \text{Bin}\left(\mu_i = \pi_i, n = 1\right) \tag{10}$$

- Remember that $\mathrm{E}(Y_i) = \pi_i$ and $\text{Var}(Y_i) = \pi_i(1 - \pi_i)$.

Logistic regression

- The linear predictor is

$$\boxed{\text{logit}(\pi_i) = \boldsymbol{x}_i^T \beta}, \tag{11}$$

- $h(\pi_i) = \text{logit}(\pi_i) = \log(\pi_i/(1-\pi_i))$ =log odds
- The variance function $v(\pi_i) = \pi_i(1-\pi_i)$ and $\phi = 1$.
- The expected value is the inverse function (logistic function)

$$\pi_i = \frac{\exp(\boldsymbol{x}_i^T \beta)}{1 + \exp(\boldsymbol{x}_i^T \beta)} \tag{12}$$

Logistic regression

- Interpretation of the parameters: β_j (except for the intercept) is the difference in logits (log odds ratio) for two subpopulations that differ on x_j by one unit.
- $\exp(\beta_j)$ (except for the intercept) is the odds ratio OR for the event for two subpopulations that differ on x_j by one unit.

Example simple logistic regression

- Simulate some data:

```
set.seed(4) #random seed
N<-30        #sample size
x<-sort(runif(N,-5,5)) #predictor
alpha<-0               #intercept
beta<-1                #slope
eta<-alpha+x*beta      #linear predictor
prob<-exp(eta)/(1+exp(eta)) #logistic function is the inverse function of the logit function
Y<-rbinom(N,size=1,prob=prob) #N samples from binomial distribution with parameters pi and n=1
miss<-sample(1:N,size=4) #missing indicator (add some real-world missing data)
Y[miss]<-NA
d.ToyLogReg<-data.frame(Y,x)
```

Example simple logistic regression

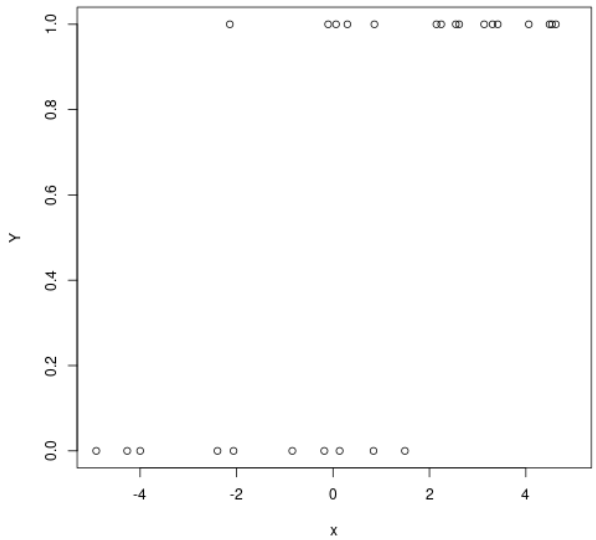
```
library(psych)
str(d.ToyLogReg)

## 'data.frame': 30 obs. of  2 variables:
##  $ Y: int  0 0 0 0 NA 1 0 0 NA 0 ...
##  $ x: num  -4.91 -4.27 -4 -2.4 -2.23 ...

headTail(d.ToyLogReg)

##        Y     x
## 1      0 -4.91
## 2      0 -4.27
## 3      0    -4
## 4      0  -2.4
## ...  ...   ...
## 27     1  4.54
## 28     1  4.62
## 29  <NA>  4.71
## 30  <NA>  4.97
```

Example simple logistic regression

```
summary(d.ToyLogReg)
```

```
##        Y                x          
##  Min.   :0.00   Min.   :-4.91  
##  1st Qu.:0.00   1st Qu.:-0.75  
##  Median :1.00   Median : 0.85  
##  Mean   :0.62   Mean   : 0.90  
##  3rd Qu.:1.00   3rd Qu.: 3.26  
##  Max.   :1.00   Max.   : 4.97  
##  NA's   :4
```

```
plot(x, Y)
```



Example simple logistic regression

Specify argument `family="binomial"`

```
m.logreg <- glm(Y ~ x, family = "binomial", data = d.ToyLogReg)
summary(m.logreg)
```

```
##
## Call:
## glm(formula = Y ~ x, family = "binomial", data = d.ToyLogReg)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.0831     0.5772    0.14     0.89
## x             0.8478     0.3306    2.56     0.01
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34.646  on 25  degrees of freedom
## Residual deviance: 19.715  on 24  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 23.71
##
## Number of Fisher Scoring iterations: 5
```

The true values are 0 for the intercept and 1 for the slope.

Wald-tests and LRT-Tests

- Tests of individual coefficients based on approximative normality are called Wald-tests with a crude assumption about the shape of the likelihood.
- The LRT takes the likelihood values as they are.
- Therefore LR-tests are usually superior to Wald-tests
- They are asymptotically equivalent.
- `confint()` constructs likelihood confidence intervals if a glm-object is given as argument.

```
m.logreg0 <- glm(Y ~ 1, family = "binomial", data = d.ToyLogReg)
anova(m.logreg0, m.logreg, test = "LRT")

## Analysis of Deviance Table
##
## Model 1: Y ~ 1
## Model 2: Y ~ x
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        25       34.6
## 2        24       19.7  1     14.9  0.00011
```

```
Stat <- 2 * (logLik(m.logreg) - logLik(m.logreg0))
as.numeric(1 - pchisq(Stat, 1))

## [1] 0.000111
```

logits and odds ratios

- `confint()` constructs "likelihood" confidence intervals

```
cbind(coef(m.logreg), confint(m.logreg))
```

```
##                    2.5 % 97.5 %
## (Intercept) 0.0831 -1.12   1.23
## x           0.8478  0.34   1.72
```

- Exponentiated coefficients: odds ratios, exp

```
cbind(exp(coef(m.logreg)), exp(confint(m.logreg)))
```

```
##                   2.5 % 97.5 %
## (Intercept) 1.09 0.325   3.42
## x           2.33 1.405   5.58
```

- Alternative with `emtrends()`: Wald intervals.

```
library(emmeans)
emtrends(m.logreg, ~1, var = "x", infer = c(TRUE, TRUE))
```

```
##  1       x.trend    SE  df asymp.LCL asymp.UCL z.ratio p.value
##  overall   0.848 0.331 Inf       0.2       1.5   2.564  0.0103
##
## Confidence level used: 0.95
```

check:

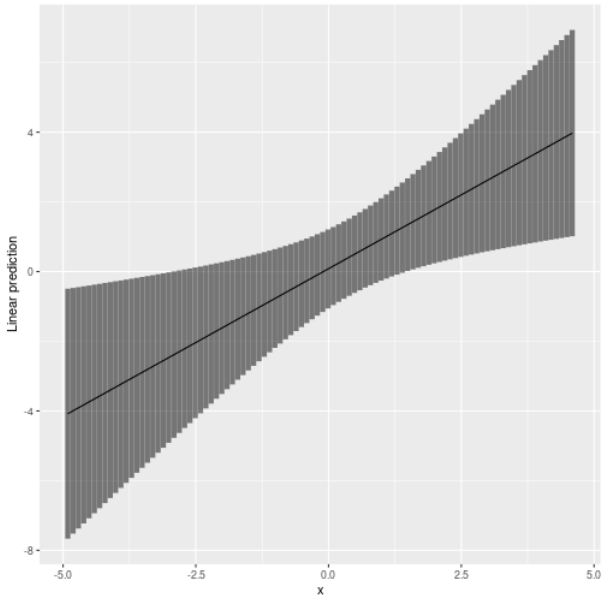
```
0.85 + c(-1, 1) * 1.96 * 0.33
```

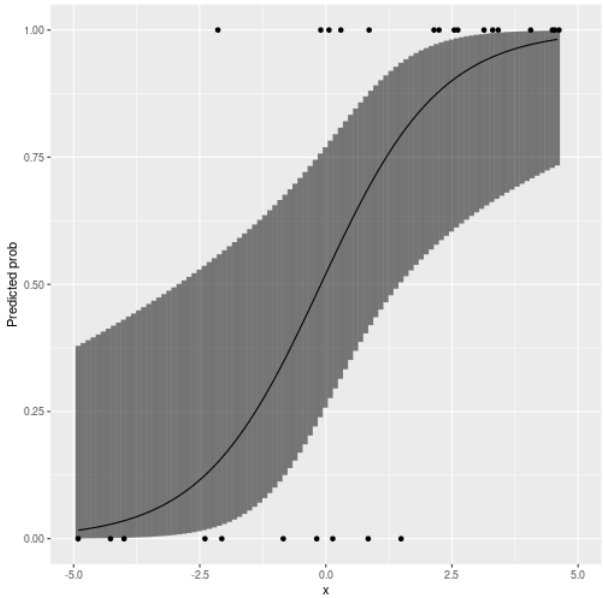
```
## [1] 0.203 1.497
```

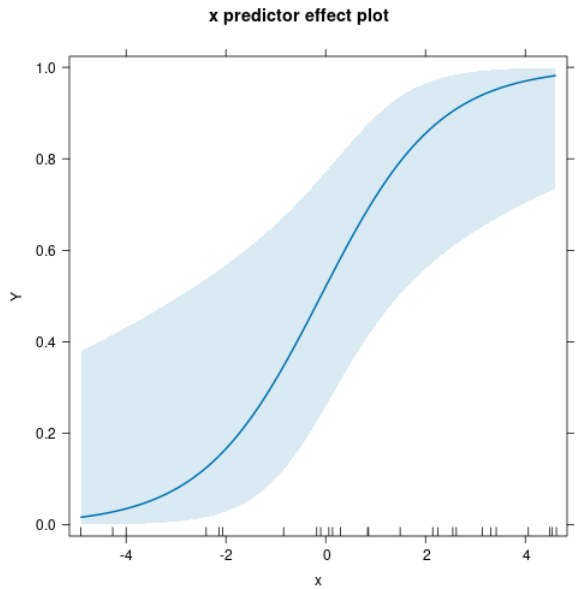
- Interpretation?

Predictions on logit and response scale

```
library(ggplot2)
emmip(m.logreg,~x,cov.reduce=function(x){seq(min(x),max(x),.1)},CIs=TRUE)
emmip(m.logreg,~x,cov.reduce=function(x){seq(min(x),max(x),.1)},type="response",CIs=TRUE)+geom_point(data=d.ToyLogReg,aes(x,as.numeric
library(effects) #alternative with effects package
plot(predictorEffects(m.logreg,"x"),axes=list(y=list(type="response"))) #alternative with effects package
```







Numerical predictions on response scale

```
emmip(m.logreg,~x,type="response",cov.reduce=function(x){seq(min(x),max(x),by=2)},CIs=TRUE,plotit=FALSE)
```

```
##      x  yvar     SE  df   LCL   UCL tvar  xvar
##  -4.91 0.017 0.0299 Inf 0.000 0.379 1    -4.91
##  -2.91 0.084 0.0941 Inf 0.008 0.501 1    -2.91
##  -0.91 0.334 0.1565 Inf 0.112 0.666 1    -0.91
##   1.09 0.732 0.1206 Inf 0.450 0.901 1     1.09
##   3.09 0.937 0.0628 Inf 0.648 0.992 1     3.09
##
## Confidence level used: 0.95
## Intervals are back-transformed from the logit scale
```

Residual analysis

What residuals are is not unambiguous:

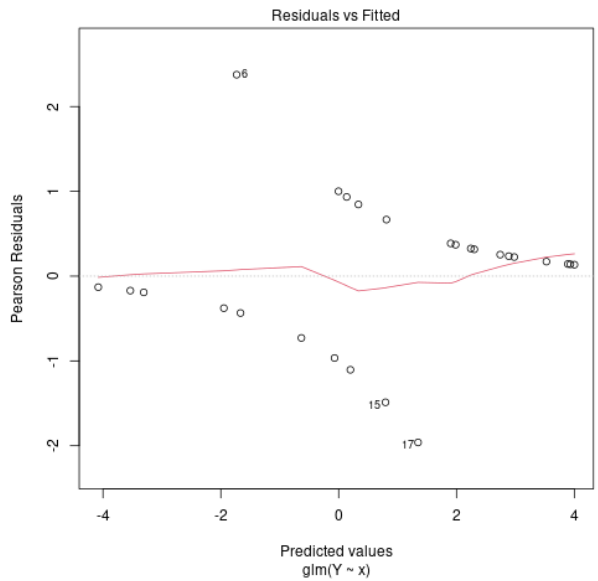
- Raw residuals (Response residuals) $R_i = Y_i - \hat{\pi}_i$
- Working residuals (transformed on the space of the linear predictor)
- Deviance residuals[2].
- Pearson residuals (Raw residuals divided by the standard deviation)

[2] $\text{sign}(Y_i - \hat{\pi}_i) \cdot \sqrt{d_i}$ with d_i as the contribution i to the deviance, would be equal to the square root of a squared residual in normal distribution.

Residual analysis*

- Working residuals against linear predictor
- Response residuals against fitted values

```
plot(m.logreg, which = 1)
```



Example 2: HIV

```
d.hiv <- read.csv("https://raw.githubusercontent.com/mcdr65/StatsRsource/master/Data/HIV.csv")
str(d.hiv)

## 'data.frame': 316 obs. of  11 variables:
##  $ id        : int  201 202 204 205 206 207 208 209 210 211 ...
##  $ age3      : int  1 1 2 1 3 1 1 2 2 1 ...
##  $ gender    : int  1 1 1 1 1 2 1 1 1 2 ...
##  $ race3     : int  4 2 5 5 5 4 4 2 2 2 ...
##  $ educ4     : int  3 4 4 3 1 4 4 3 3 1 ...
##  $ employment: int  0 0 1 1 0 0 0 0 0 0 ...
##  $ disability: int  0 1 0 0 1 1 1 1 1 1 ...
##  $ dep       : int  0 0 1 0 1 1 1 1 0 1 ...
##  $ anxpoms8  : int  NA 0 1 1 1 0 1 1 0 1 ...
##  $ paindic   : int  1 1 1 0 1 1 0 1 0 1 ...
##  $ aids      : int  1 0 0 0 1 1 0 1 0 0 ...

isafactor <- c(1:11)
d.hiv[, isafactor] <- lapply(d.hiv[, isafactor], as.factor)
levels(d.hiv$age3) <- c("<39", "40-49", ">50")
levels(d.hiv$gender) <- c("male", "female", "transgender")
levels(d.hiv$race3) <- c("black", "white", "mix")
levels(d.hiv$employment) <- c("no", "yes")
levels(d.hiv$disability) <- c("no", "yes")
levels(d.hiv$dep) <- c("no", "yes")
levels(d.hiv$paindic) <- c("no", "yes")
levels(d.hiv$aids) <- c("no", "yes")
```

Example 2: HIV

```
str(d.hiv)

## 'data.frame': 316 obs. of  11 variables:
##  $ id         : Factor w/ 316 levels "201","202","204",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ age3       : Factor w/ 3 levels "<39","40-49",..: 1 1 2 1 3 1 1 2 2 1 ...
##  $ gender     : Factor w/ 3 levels "male","female",..: 1 1 1 1 1 2 1 1 1 2 ...
##  $ race3      : Factor w/ 3 levels "black","white",..: 2 1 3 3 3 2 2 1 1 1 ...
##  $ educ4      : Factor w/ 4 levels "1","2","3","4": 3 4 4 3 1 4 4 3 3 1 ...
##  $ employment: Factor w/ 2 levels "no","yes": 1 1 2 2 1 1 1 1 1 1 ...
##  $ disability: Factor w/ 2 levels "no","yes": 1 2 1 1 2 2 2 2 2 2 ...
##  $ dep        : Factor w/ 2 levels "no","yes": 1 1 2 1 2 2 2 2 1 2 ...
##  $ anxpoms8   : Factor w/ 2 levels "0","1": NA 1 2 2 2 1 2 2 1 2 ...
##  $ paindic    : Factor w/ 2 levels "no","yes": 2 2 2 1 2 2 1 2 1 2 ...
##  $ aids       : Factor w/ 2 levels "no","yes": 2 1 1 1 2 2 1 2 1 1 ...
```

Example 2: Logistic regression

In the summary, we see marginal Wald tests (based on approximative normality).

```
m.1 <- glm(aids ~ age3 * gender + race3, family = "binomial", data = d.hiv)
m.1b <- glm(aids ~ age3 * gender, family = "binomial", data = d.hiv)
m.1c <- glm(aids ~ age3 + gender, family = "binomial", data = d.hiv)
m.0 <- glm(aids ~ 1, family = "binomial", data = d.hiv)
summary(m.1b)
```

```
##
## Call:
## glm(formula = aids ~ age3 * gender, family = "binomial", data = d.hiv)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -0.164      0.257   -0.64    0.523
## age340-49                      0.598      0.336    1.78    0.075
## age3>50                        0.532      0.359    1.48    0.138
## genderfemale                   0.164      0.562    0.29    0.770
## gendertransgender              0.387      0.718    0.54    0.590
## age340-49:genderfemale        -0.598      0.684   -0.87    0.382
## age3>50:genderfemale          -0.974      0.749   -1.30    0.194
## age340-49:gendertransgender  -16.387    594.164   -0.03    0.978
## age3>50:gendertransgender     -1.266      1.055   -1.20    0.230
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 437.45  on 315  degrees of freedom
## Residual deviance: 421.35  on 307  degrees of freedom
## AIC: 439.3
##
## Number of Fisher Scoring iterations: 14
```

Example 2: Sequential LR tests

```
anova(m.1b, test = "LRT")

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: aids
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                          315        437
## age3         2     1.10       313        436    0.578
## gender       2     4.94       311        431    0.084
## age3:gender  4    10.06       307        421    0.039
```

- One could proceed with different model comparisons.

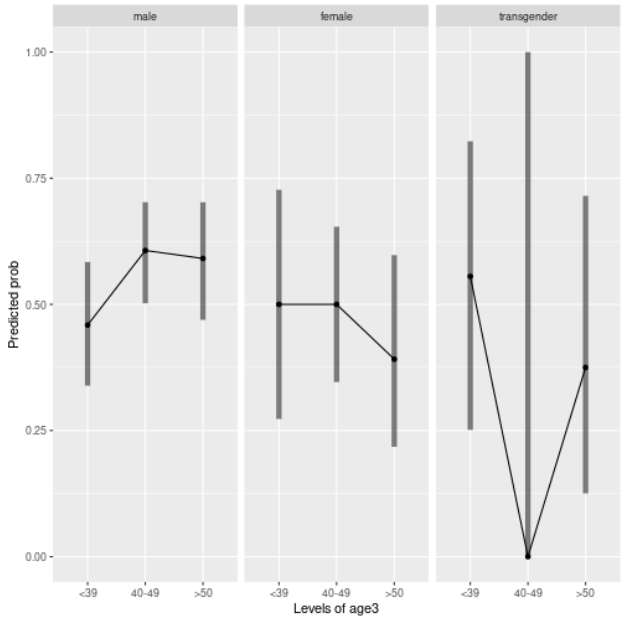
Example 2: Marginal Tests

```
drop1(m.1b, test = "LRT")

## Single term deletions
##
## Model:
## aids ~ age3 * gender
##             Df Deviance AIC  LRT Pr(>Chi)
## <none>              421 439
## age3:gender  4      431 441 10.1    0.039
```

Predictions

Different effects can be visualized with `emmeans::emmip`, on the scale of the linear predictor or on the response scale.

```
emmip(m.1b, ~age3 | gender, type = "response", CIs = TRUE)
```



Collapsibility of effect measures

- Given: Binary treatment indicator X and continuous C uncorrelated with X.
- Question: Does the effect of X change when we condition on non-confounding C?
- We know this is not the case for linear models.

Collapsibility in linear models

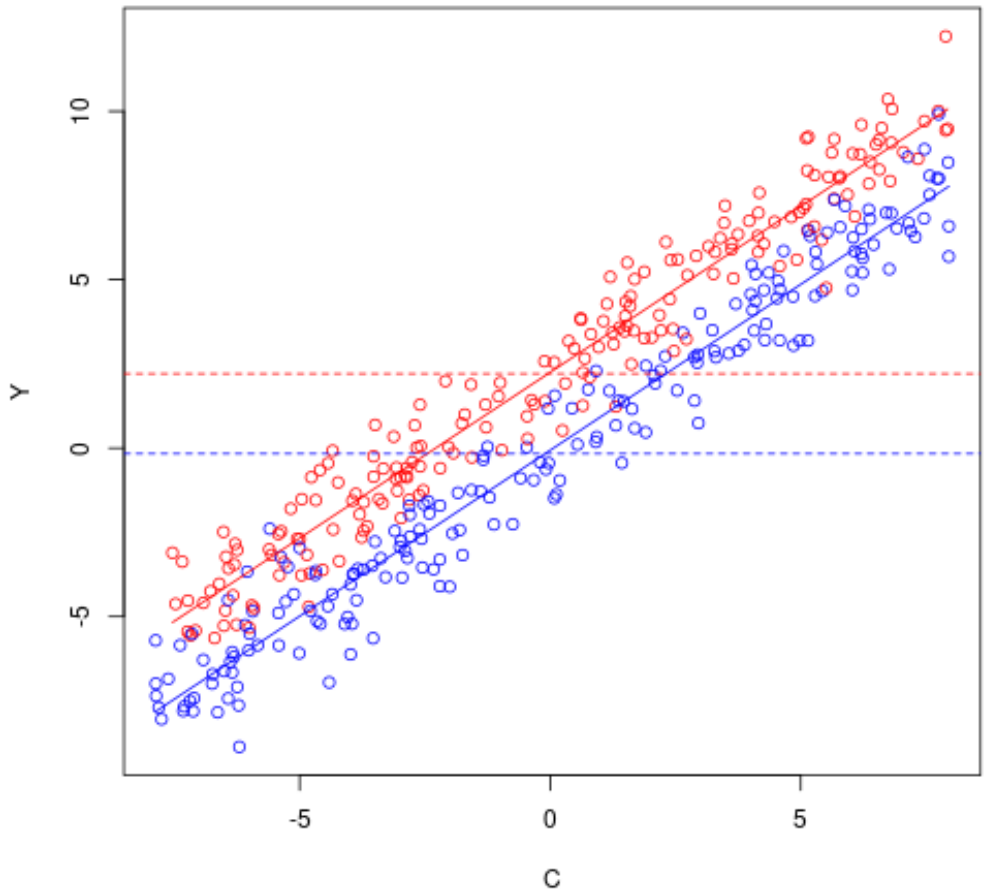


Figure: In linear models, marginal (dotted) and conditional group effects are equal in the absence of confounding.

Collapsibility in linear models

```
modlinM  #marginal

##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)           XB
##      -0.164        2.375

modlinC  #conditional

##
## Call:
## lm(formula = Y ~ X + C)
##
## Coefficients:
## (Intercept)           XB            C
##     -0.0631       2.3152       0.9850
```

Noncollapsibility of the odds ratio

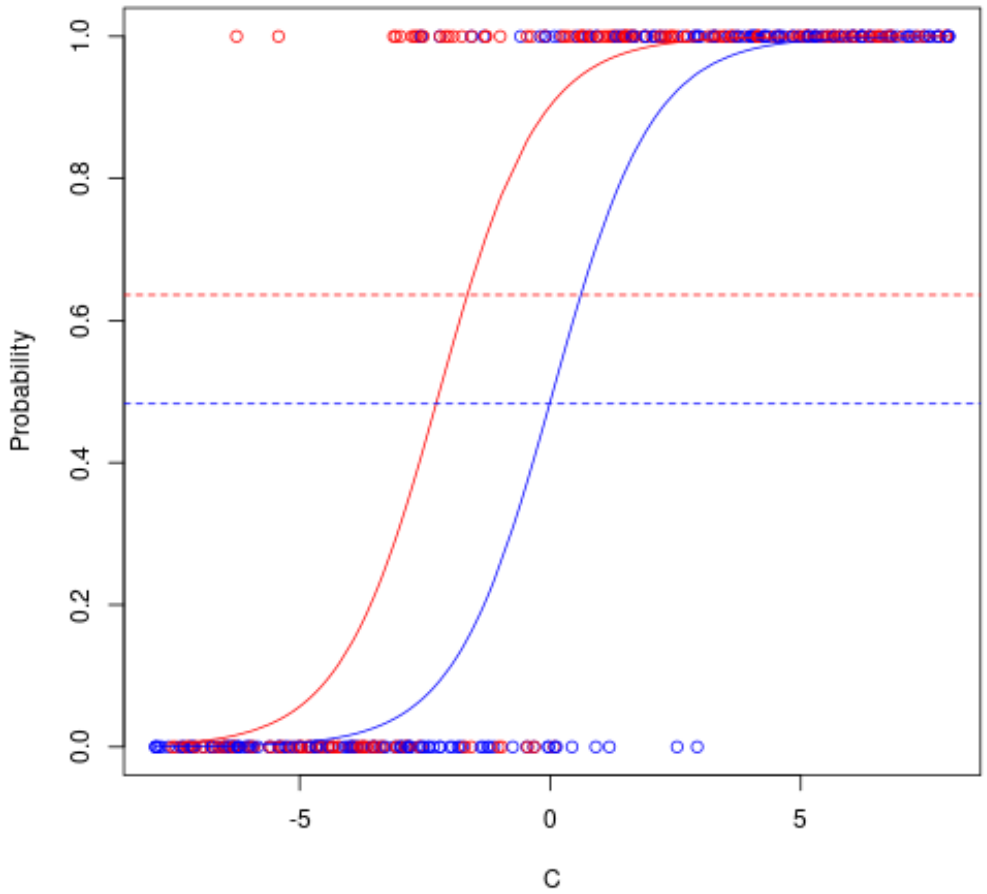


Figure: In logistic models, marginal (dotted) and conditional (on C) group effects differ even in the absence of confounding.

Noncollapsibility of the odds ratio

The marginal OR is always shifted toward the null compared to the conditional OR!

```
modM  #marginal

##
## Call:  glm(formula = Ydich ~ X, family = "binomial")
##
## Coefficients:
## (Intercept)           XB
##      -0.080        0.612
##
## Degrees of Freedom: 399 Total (i.e. Null);  398 Residual
## Null Deviance:     550
## Residual Deviance: 541  AIC: 545

modC  #conditional

##
## Call:  glm(formula = Ydich ~ X + C, family = "binomial")
##
## Coefficients:
## (Intercept)           XB            C
##     -0.0563       2.2816       1.0072
##
## Degrees of Freedom: 399 Total (i.e. Null);  397 Residual
## Null Deviance:     550
## Residual Deviance: 152  AIC: 158
```

Exercise: Reproduce (approximately) the point estimates using the plot on the former slide!

Noncollapsibility of the odds ratio

- When the expected probability of outcome is modeled as a nonlinear function of the exposure, the marginal effect cannot be expressed as a weighted average of the conditional effects[3].
- In the absence of confounding or when confounding is adjusted appropriately, both the marginal OR and conditional OR are valid measures.
- They are unbiased estimators for two different parameters.
- Report the marginal OR if the average effect at the population level is of interest.
- Report the conditional OR if the conditional effect at the individual or subgroup level is of interest.

[3] Jensens inequality provides theoretical justification for this noncollapsibility in the absence of confounding, requiring that the marginal OR is always shifted toward the null compared to the conditional OR.