Experimental Design and Data Analysis, Lecture 10

Eduard Belitser

VU Amsterdam

Lecture overview

1. generalized linear models
   - logistic regression
   - Poisson regression

generalized linear models

Setting

An experiment with:

- an outcome Y that is has a different nature than in ANOVA or linear regression;
- one or more numerical explanatory variables $X_1, \ldots, X_p$.
- one or more factor explanatory variables. ("independent variable").

The purpose is to explain Y by a linear function of X.

EXAMPLE Educational study with outcome passed the exam or not and explanatory variable number of pupils per teacher. Y is binary.

EXAMPLE The number of plant species on a Galapagos Island, with explanatory variables area, highest elevation, distance to nearest island, distance to Santa Cruz island and area of adjancent island. Y is a count.

EXAMPLE Political study with outcome party identification with explanatory variables age, education level and income. Y is multinomial (categorical).

doesn't have to be a realization of norm. r.v. → obs. can be as realiz. of some other dist.

dist. doesn't have to be normal, it can be binom., exp., Poisson etc.

2 poss. val.s = 0 or 1

ANOVA, ANCOVA etc. wouldn't be a good model

model w/ Poisson dist.

design model s.t. dist. of obs. dep. on param.s and those param.s dep. on these var.s = predict.s

extension of binom. dist.

Different models

For each of the three examples a different model applies.

- For binary responses, the logistic regression model assumes:

$$\log \frac{P(Y=1)}{P(Y=0)} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$

prob. of succ. $P(Y=1) = p$, $P(Y=0) = 1-p$ prob. of fail

lin. comb. of var.s

odds ratio = binom. dist. r.v. (b/c r.v. takes 2 val.s = 0 and 1)

- For multinomial responses, the multinomial logit model assumes:

$$\log \frac{P(Y=C_i)}{P(Y=C_1)} = \beta_0^i + \beta_1^i X_1 + \ldots + \beta_p^i X_p,$$

lin. comb. of pred.s

★ instead of 0 and 1, you have i poss. responses

where C_1 is the reference class of the categorical responses.

- For count responses, the Poisson regression model assumes:

$$\log E(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$

λ

lin. comb. of pred.s

we assume that cert. r.v. has Poisson dist., all we have to spec. param λ = expect. of Poisson r.v.

logistic regression

Setting

An experiment with:

- an outcome Y that is 0 or 1 ("binary dependent variable");
- one or more numerical explanatory variables $X_1, \ldots, X_p$.
- one or more factor explanatory variables $F_1, \ldots, F_m$.

just like ANCOVA

The purpose is to explain Y by a function of X's and F's .

EXAMPLE A subject participates or not in an internet survey presented in 3 formats at 3 different days of the week.

EXAMPLE Educational study with outcome passed the exam or not and explanatory variable number of pupils per teacher.

EXAMPLE Medical study with outcome patient died or not with explanatory variables type of treatment, sex and age.

Design

Logistic regression can be used for factorial experiments, in a regression setting, for ANCOVA, and for experiments with blocks.

- The design is the same as for the corresponding experiment.

Logistic regression is also used in a case-control setting.

- Consider a population consisting of 2 subpopulations of units with outcome 0 and with outcome 1, respectively ("controls" and "cases").
- Independently choose random samples of units from the two subpopulations.
- Measure the explanatory variables for these units.

The case-control design has the advantage that the numbers of cases and controls in the samples can be fixed in advance (and made approx. equal).

Logistic regression model

- Response Y is categorical (0-1) $\rightarrow$ cannot use lin.regr./anova/ancova.
- In this case, we model $\Pr(Y = 1)$ as a function of explanatory variables.
- The logistic regression model assumes that outcome $Y_k \in \{0, 1\}$ satisfies

$$P(Y_k = 1) = \Psi(\boldsymbol{x}_k^T \theta) = \frac{1}{1 + e^{-\boldsymbol{x}_k^T \theta}}, \qquad P(Y_k = 0) = 1 - P(Y_k = 1),$$

prob. of succ. — lin. comb.

$\boldsymbol{x}_k^T \theta = \mu + \alpha_{f(k)} + \ldots + \beta_1 x_{k1} + \ldots$, $f(k) \in \{1, \ldots, I\}$ is the factor level of observation Y_k, $\boldsymbol{x}_k = (1, \ldots, 0, 1, 0, \ldots, x_{k1}, \ldots)^T$ is the k-th vector of predictor values, $\theta = (\mu, \alpha_1, \ldots, \beta_1, \ldots)^T$ is the parameter vector.

everything which is in model here

- $\Psi(x) = 1/(1 + e^{-x})$, $\Psi : \mathbb{R} \mapsto [0, 1]$, is called logistic function. — always b.w. 0 and 1
- The explanatory variables can be either numerical or categorical, or a mix.
- As in lin.regr./anova/ancova, we can test for factors/variables, their interactions, estimate the parameters, and predict future observations.
- In R: `glm(y~f1+...+x1+...,family=binomial,data=mydata)`

nothing changes w.r.t. model formula

If the categorical response variable has more than 2 values, one extends the usual logistic regression to multinomial logistic regression (implem. in R by special packages).

Example: logistic regression with one factor and one contin. predictor

- For example, for a single factor with I levels and a single numerical explanatory variable the logistic regression model assumes that the outcome Y_{ik} of a unit measured at level i of the factor and having explanatory variable X_{ik} satisfies

$$P(Y_{ik} = 1) = \Psi\left(\mu + \alpha_i + \beta X_{ik}\right), \quad i = 1, \ldots, I, \quad k = 1, \ldots, N,$$

- Want to tests the hypotheses $H_0 : \alpha_1 = \ldots = \alpha_I = 0$, and $H_0 : \beta = 0$, i.e., the factor and/or explanatory variable do not influence the outcome.
- Also estimate the factor effects $\alpha_1, \ldots, \alpha_I$ and the regression parameter β.

The outcome Y is like a coin-toss; the probability $P(Y = 1)$ of "heads" is modelled. The linear predictor $\mu + \alpha_i + \beta X_{ik}$ can take any real value. The logistic function maps this into a probability: a number between 0 and 1. A bigger linear predictor gives a probability of heads closer to 1.

Logistic regression: odds

- The odds is $o = \frac{P(Y=1)}{P(Y=0)}$. This means that the probability of "success" $P(Y=1)$ is o times as big as the probability of "failure" $P(Y=0)$.
- This is a linear model for the log odds: $\log o_k = \boldsymbol{x}_k^T \theta$ or $o_k = e^{\boldsymbol{x}_k^T \theta}$.
- For example, for the logistic regression with one factor and one contin. predictor,

$$o_{ik} = \frac{P(Y_{ik}=1)}{P(Y_{ik}=0)} = e^{\mu+\alpha_i+\beta X_{ik}}, \quad \text{or} \quad \log o_{ik} = \mu + \alpha_i + \beta X_{ik}.$$

- A change Δ in the linear predictor $\mu + \alpha_i + \beta X_{ik}$ multiplies the odds by e^{Δ}. For example,
  - an increase of predictor X by one unit multiplies the odds by e^{β}.
  - a change from level i to level i' multiplies the odds by $e^{\alpha_{i'} - \alpha_i}$.

Analysis in R: data input, graphics

In the data set esoph.txt, the column cancer indicates whether the individual (1–1175) suffers from cancer of the esophagus (gullet). The first three columns give the age rounded to a multiple of 10, alcohol consumption, and tobacco use.

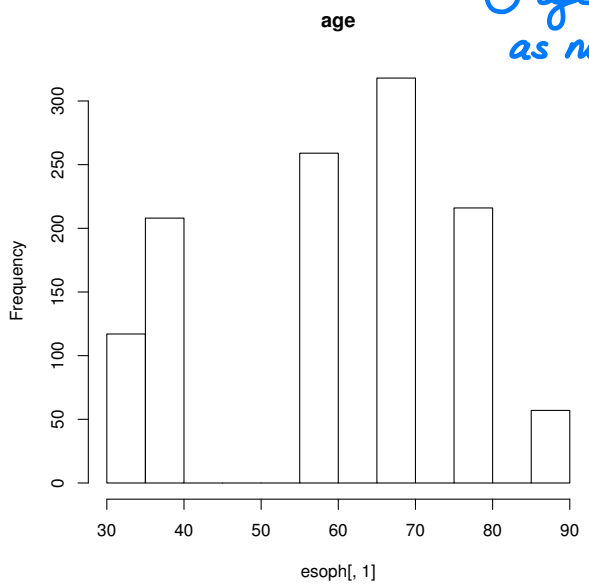
```
> esoph=read.table("esoph.txt",h=T)
> esoph
     age alc tob cancer
1     30  20   5      0
2     30  20   5      0
3     30  20   5      0
[ a lot of output deleted ]
1173  90 140   5      0
1174  90 140  15      1
1175  90 140  15      0
```

y (response var.)

0 → no cancer, 1 = has cancer

```
> hist(esoph[,1],main="age")
```

⊙ age is treat. as num.



The histogram shows the age distribution.

Analysis in R: summary

use to see counts = how many obs in cert. cells

```
> tot=xtabs(~alc+tob,data=esoph)
> tot
     tob
alc    5  15  25  35
  20 270  94  47  33
  60 213 102  77  38
  100 80  68  22  19
  140 40  30  19  23
```

you get counts w.r.t. cells alc. and tob.

The table shows the total numbers of individuals for each combination of levels of alcohol and tobacco use.

added to compute # of ppl. whose label cancer = 1

```
> tot.c=xtabs(cancer~alc+tob,data=esoph)
> round (tot.c/tot,2)
     tob
alc    5   15   25   35
  20 0.03 0.11 0.11 0.15
  60 0.16 0.17 0.19 0.24
  100 0.24 0.28 0.27 0.37
  140 0.40 0.40 0.37 0.43
```

most of alc. and most of tob.

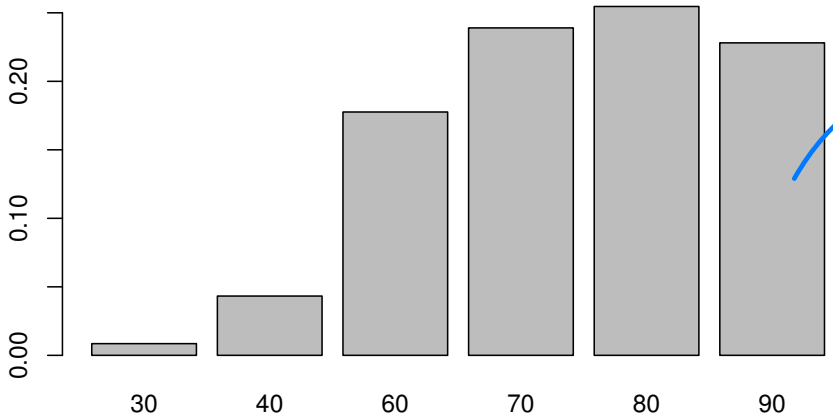
biggest frac. of having cancer

The table shows the percentage of individuals with cancer for every combination of levels of alcohol and tobacco use.

Analysis in R: graphics

```
> totage=xtabs(~age,data=esoph)
> barplot(xtabs(cancer~age,data=esoph)/totage)
```

how ppl w/cancer are dist. acc. to age



you look at age as val. of func. and prob. as argument → you might think it's quad. func. → then you might take age^2 as new var.

The barplot shows the percentage per age-group. Since it doesn't look very linear, we will add age^2 as an explanatory variable in the next slide.

Remark. This is just to demonstrate that one can create and include other variable(s) in the model, this is not necessarily good thing to do, for example the variable age^2 is not well interpretable and, besides, it will turn out to be not useful.

Analysis in R: estimation and testing

```
> esoph$age2=esoph$age^2
> esophglm=glm(cancer~age+age2+alc+tob,data=esoph,family=binomial)
> summary(esophglm)

             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.8072283  1.5850673  -6.187 6.12e-10 ***
age          0.1688542  0.0491991   3.432 0.000599 ***
age2        -0.0009608  0.0003776  -2.545 0.010934 *
alc          0.0162614  0.0021092   7.710 1.26e-14 ***
tob          0.0256080  0.0081412   3.145 0.001658 **
```

age² → new col.

$\hat{\mu}$ β_1 β_2 β_3 β_4

$\beta_1 \neq 0$ → in model

in model

prob. val.s for testing the hypot. that corresp. β is equal to 0.

sig.

The R-function `glm` (generalized linear model) is used instead of `lm` to create the glm object. The option `family=binomial` overrules the default normal model (which gives `lm`). The 4 explanatory variables are inserted here as numerical. The estimated odds is $\hat{o}_k = \frac{\widehat{P(Y_k=1)}}{\widehat{P(Y_k=0)}} \approx \exp\{-9.8 + 0.17\text{age}_k - 0.00096\text{age2}_k + 0.016\text{alc}_k + 0.026\text{tob}_k\}$. The positive signs of the parameter estimates mean that higher values of these variables give higher probability of cancer. For instance, raising tobacco by 1 increases the linear predictor by 0.0256080 and increases the odds of cancer by a factor $e^{0.0256080} = 1.026$. For age the dependence is parabolic: from 25 to 30 years the odds increase by $\exp\{0.17 \cdot (30 - 25) - 0.00096 \cdot (30^2 - 25^2)\} = 1.786932$.

compute est. odds, we subst. corresp. val.s

b/c of age²

approx. twice increase in odds

Analysis in R: `glm` instead of `lm`

- Once a `glm` object is created one can access the various components of the results in the same way as for any other linear model R-object, using functions such as `summary`, `anova`, `drop1`, `coef`, `residuals`, etc.
- For example, `mod=glm(y~x1+x2,data,family=binomial)`, and the command `summary(mod)` displays the (MLE) estimates of the model coefficients and individual tests that these coefficients are zero.
- Pay attention to the parametrization (in case of factors) and to the order of the variables in the model formula. Need to specify the test (for GLM's, `"Chisq"`) in testing commands, e.g, `drop1(mod,test="Chisq")`.
- Instead of anova table, `anova(mod,test="Chisq")` yields the so called deviance tables, which are used to examine the progressive fit of the model as each covariate/factor is added to the model.
- The safest way (and to have the full control of what you test) is to use `anova(mod1,mod2,test="Chisq")` or `drop1(mod,test="Chisq")`.
- Diagnostics for GLM's is not as straightforward as for linear models, and will not be treated in this course. For example, there are at least 5 types of residuals and 2 types of fitted values for GLM's ($\hat{\mu}_k$ and $x_k^T\hat{\theta}$).

def. param. is treat.

gets all relevant p-val.s of fac.s

not F-test b/c it's not norm.ly dist.

involves some kind of approx. = chi-sq.

mod1 = small mod, mod2 = big model

has smth we test for

Analysis in R: estimation and testing (1)

```
> esoph$age=factor(esoph$age); esoph$alc=factor(esoph$alc)
> esoph$tob=factor(esoph$tob) # note: the variables are factors now
> glm2=glm(cancer~age+alc+tob,data=esoph,family=binomial); summary(glm2)
[ some output deleted ]
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9108     1.0302  -5.738 9.59e-09 ***
age40         1.6095     1.0675   1.508 0.131631
age60         2.9752     1.0242   2.905 0.003673 **
age70         3.3584     1.0198   3.293 0.000991 ***
age80         3.7270     1.0252   3.635 0.000278 ***
age90         3.6818     1.0644   3.459 0.000542 ***
alc60         1.1216     0.2384   4.704 2.55e-06 ***
alc100        1.4471     0.2628   5.506 3.68e-08 ***
alc140        2.1154     0.2876   7.356 1.90e-13 ***
tob15         0.3407     0.2054   1.659 0.097159 .
tob25         0.3962     0.2456   1.613 0.106708
tob35         0.8677     0.2765   3.138 0.001701 **
```

we made # of coeff.s bigger = now each level becomes var. in model

μ α_2 α_3 β_2 β_3

not signif.

not signif.

⊙ you don't see age30, alc60 b/c of treat. param.

you have to spec. group which you comp. est. odds

In the previous model, tob, alc and age were numeric, here they are categorical, treated as factor. The variable age2 is dropped. For example, the estimated odds for the group (age70, alc20, tob35) is $\hat{o} \approx \exp\{-5.91 + 3.36 + 0 + 0.87\} = e^{-1.68}$.

odd. decr.

$\beta_1 = 0$

Analysis in R: estimation and testing (2)

Recall that $\widehat{P(Y_k = 1)} = \Psi(\boldsymbol{x}_k^T \hat{\theta})$. For example, the estimate of the probability of cancer for the group (age70, alc20, tob35) is computed as $\Psi(\text{Intercept} + \text{age70} + \text{alc20} + \text{tob35}) = 0.1564698$.

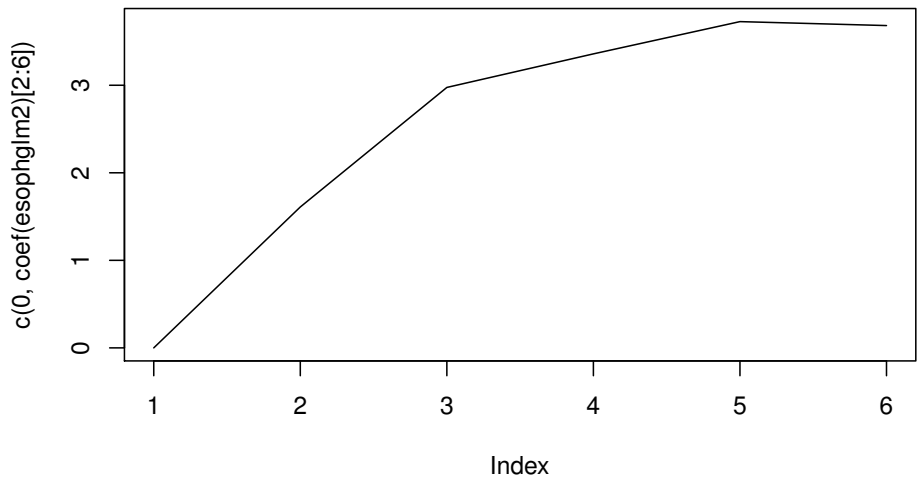
prob. of cancer = relat. low

In R, all $\widehat{P(Y_k = 1)}$ are obtained by `fitted(glm2)`. To predict the probability of cancer for `newdata`, use `predict(glm2,newdata,type="response")`, for example, `newdata=data.frame(age="70",alc="20",tob="35")`.

you can take any combination of levels

Make a graph of the coefficients for the different age categories:

```
> plot(c(0,coef(glm2)[2:6]),type="l")
```



By inserting the variables as factors each level gets its own parameter, and we can look at the dependence on levels. Disadvantage: (too) many parameters.

Analysis in R: estimation and testing (3)

for glms, you always have to spec. test

```
> drop1(glm2,test="Chisq")
Single term deletions
       Df Deviance    AIC    LRT   Pr(Chi)
<none>      898.86 922.86
age     5   976.37 990.37 77.511 2.782e-15 ***
alc     3   964.91 982.91 66.054 2.984e-14 ***
tob     3   909.46 927.46 10.599   0.01411 *
```

all are signif.

we are looking at fac.s as a whole

As the variables are factors now, the drop1 command reduces the list of the p-values to one p-value per variable in the model formula, for testing the null hypothesis that the factor has no effect. All three factors are significant. The anova command works too, but gives "sequential" tests, which are hard to interpret (only the last p-value can be well interpreted). Another (and the best) way to get correct p-values, for example, for the factor alc: glm3=glm(cancer~age+tob,data=esoph,family=binomial), then anova(glm3,glm2) will give the right p-values for the factor alc.

testing for smth absent in small model but present in big model

Aggregated data format (for logistic model)

- Measurements with the same values of all explanatory variables need not be represented by separate lines in the data matrix.
- Instead we can count for every combination of explanatory variables the total numbers of 0's and 1's.
- One line in dataset `esophshort.txt` contains the aggregated data of lines with equal values of the explanatory variables (factors) in the dataset.

to reduce data w/out losing info

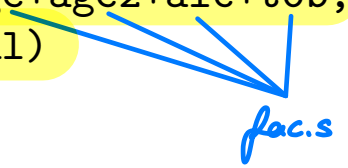
```
> esophshort=read.table("esophshort.txt",header=TRUE)
> esophshort$age2=esophshort$age^2
> head(esophshort)
```

cancer → ncases; no cancer → ncontrols

	age	alc	tob	ncases	ncontrols
1	30	20	5	0	40
2	30	20	15	0	10
3	30	20	25	0	6
4	30	20	35	0	5
5	30	60	5	0	27
6	30	60	15	0	7

Aggregated data format (2)


```
> shortglm=glm(cbind(ncases,ncontrols)~age+age2+alc+tob,
+              data=esophshort,family=binomial)
> summary(shortglm)
[ some output deleted ]
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.8072283  1.5850903  -6.187 6.13e-10 ***
age          0.1688542  0.0491997   3.432 0.000599 ***
age2        -0.0009608  0.0003776  -2.545 0.010935 *
alc          0.0162614  0.0021092   7.710 1.26e-14 ***
tob          0.0256080  0.0081413   3.145 0.001658 **
```



The output is identical to that of the earlier analysis with the "long" data, using the explanatory variables as `numeric` variables.

This aggregated format in the form of pair `(success,failure)`, the counts of successes and failures for each combination of levels of the factors (or values of numeric variables), is one of 3 possible ways to specify the responses in `R` for the logistic model. This format is not useful if there is a continuous predictor in the model, taking different values for different individuals (e.g., diff. ages for diff. individuals).

Testing interaction between factor and contin. predictor (1)

Condsider a model with one factor alc and one contin. predictor age.

```
> esoph$age=as.numeric(esoph$age)
> glm3=glm(cancer~age+alc,data=esoph,family=binomial)
      Df Deviance     AIC    LRT  Pr(>Chi)
<none>      925.23  935.23
age    1   983.67  991.67 58.440 2.096e-14 ***
alc    3  1012.48 1016.48 87.244 < 2.2e-16 ***
```

not fac., it's num. var.

only this p-val. is relevant.

b/c order matters

Recall the model we are actually studying

$$P(Y_{in} = 1) = \Psi(\mu + \alpha_i + \beta X_{in}) = 1/\left(1 + e^{-(\mu+\alpha_i+\beta X_{in})}\right),$$

both the factor and contin. predictor are in the model (as in ancova).

However, the coefficient(s) β (reflecting the influence of the continuous predictor) may depend on the level of the factor, i.e.,

$$P(Y_{in} = 1) = \Psi(\mu + \alpha_i + \beta_i X_{in}) = 1/\left(1 + e^{-(\mu+\alpha_i+\beta_i X_{in})}\right).$$

In this case we say that the corresponding factor and variable interact.

Testing interaction between factor and contin. predictor (2)

Testing for no interaction between the factor and predictor:

$H_0 : \beta_1 = \ldots = \beta_I.$ → all βs are same.

In R, to test for interaction (in logistic model and ANCOVA) between factor and contin. predictor, simply include the interaction term in the model formula, e.g., y∼f+x+f:x or y∼f*x.

Testing for the interaction between factor alc and predictor age:

interact. b.w. → always at end

```
> glm4=glm(cancer~age*alc,data=esoph,family=binomial)
> anova(glm4,test="Chisq") # only the last p-value is relevant
[  some output deleted ]
          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       1174    1072.13
age        1   59.647       1173    1012.48 1.135e-14 ***
alc        3   87.244       1170     925.23 < 2.2e-16 ***
age:alc    3    4.549       1167     920.68     0.208
```

only relevant p-val.

Only the last p-value is relevant which always concerns interaction for models with interaction. We conclude that $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$ is not rejected, i.e., there is no interaction between factor alc and predictor age.

Testing for interaction between factors and contin. variables in ANCOVA is the same.

From logistic regression to machine learning prediction

- Fitting the observed data $(X_1, Y_1), \ldots, (X_N, Y_N)$ in logistic regression

feature

obs. binom r.v. = yes/no

$$P(Y_k = 1) = \frac{1}{1 + e^{-x_k^T \theta}}, \quad k = 1, \ldots, N,$$

prob. of label = yes

= we learn our model

we obtain (by the maximum likelihood) an estimate $\hat{\theta}$ of the parameter θ.

- For a new predictor vector X_{new}, we can predict its success probability

$$\hat{P}_{new} = \frac{1}{1 + e^{-x_{new}^T \hat{\theta}}}.$$

to classify yes/no

we look at its features and by using what we learn, we compute this prob.

- Now use $\hat{P}_{new}$ to predict the new label $\hat{Y}_{new}$ as

$$\hat{Y}_{new} = \begin{cases} 1, & \text{if } \hat{P}_{new} \geq p_0 \\ 0, & \text{if } \hat{P}_{new} < p_0 \end{cases}$$

for some threshold $p_0 \in [0, 1]$.

when you obs. data label. yes/no & features = var.s

- This yields one of the commonly used prediction methods in machine learning, which you may have had in one of you machine learning courses.

Poisson regression

Setting and design

An experiment with:

- an outcome Y that is a count;
- one or more numerical explanatory variables $X_1, \ldots, X_p$.
- one or more factor explanatory variables. ("independent variable").

count of events = how many times it happened

usually rare events

The purpose is to explain Y by a function of X.

count

EXAMPLE The number of plant species on a Galapagos Island, with explanatory variables area, highest elevation, distance to nearest island, distance to Santa Cruz island and area of adjancent island.

EXAMPLE The number of military coups in some countries with explanatory variables number of years country ruled by military oligarchy, number of political parties and population size.

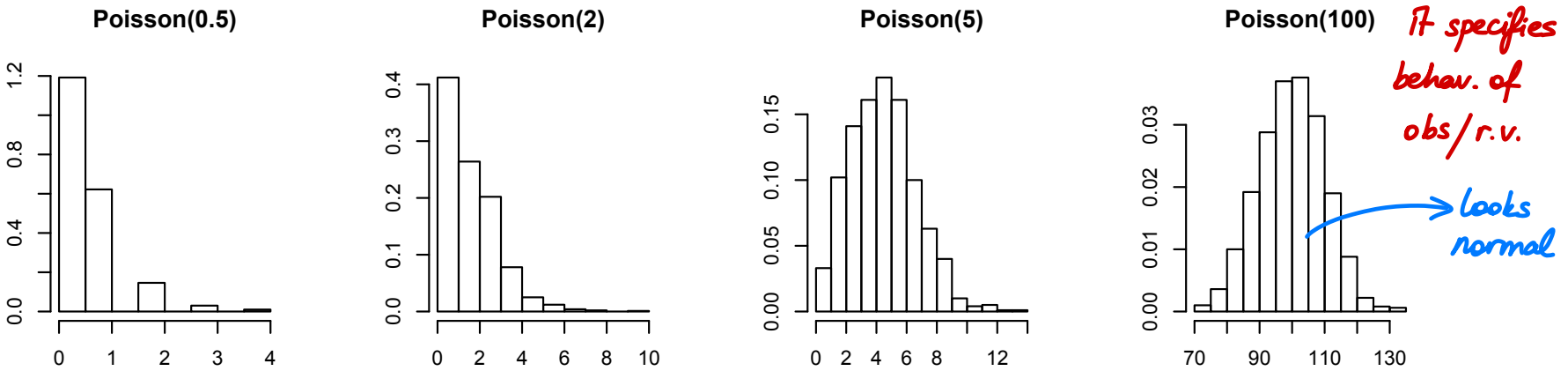
Design. Poisson regression can be used for factorial experiments, in a regression setting, for ANCOVA, and for experiments with blocks. The design is the same as for the corresponding experiment.

The Poisson distribution

- A random variable Y is said to have the Poisson(λ)-distribution, $\lambda > 0$, if

$$P(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots.$$

- If $Y \sim \text{Poisson}(\lambda)$, then $E(Y) = \text{Var}(Y) = \lambda$.
- Hence, the larger the parameter, the larger the values of Y on average and the larger the spread in the values of Y.
- For very large λ, the Poisson(λ)-distribution is approximately equal to a normal distribution with mean $\mu = \lambda$ and variance $\sigma^2 = \lambda$.



Analysis

- In Poisson-regression, the parameter λ is modelled as:

$$\log \lambda = \mu + \alpha_i + \ldots + \beta_1 X_1 + \ldots, \quad \text{or} \quad \lambda = e^{\mu+\alpha_i+\ldots+\beta_1 X_1+\ldots},$$

always pos. b/c λ is always pos.

on the right: the combination of (numerical and/or categorical) variables.

- For each Y_k the parameter λ_k is modelled differently, since the values of involved factors/predictors will differ for diff. observations: $\lambda_k = e^{x_k^T \theta}$.
- For example, for the Poisson regression with one factor (with I levels) and one continuous predictor X,

$$Y_{in} \sim \text{Poisson}(\lambda_{in}), \quad \lambda_{in} = e^{\mu+\alpha_i+\beta X_{in}}, \quad i = 1, \ldots, I, \ n = 1, \ldots, N.$$

$\hat{\lambda}$

obs. val.

- Hence, the variances are different as well. This means that the response residuals $Y_{in} - \hat{Y}_{in} = Y_{in} - e^{\hat{\mu}+\hat{\alpha}_i+\hat{\beta} X_{in}}$ are not from one fixed distribution, hence a normal QQ-plot of these response residuals is not relevant!

model checks don't work here

Instead, the deviance residuals are useful for diagnostic plots. Deviance is a measure of the discrepancy between the "full model" and the model under consideration. Deviance residuals are response residuals scaled by the deviance of that observation.

Analysis in R: data input

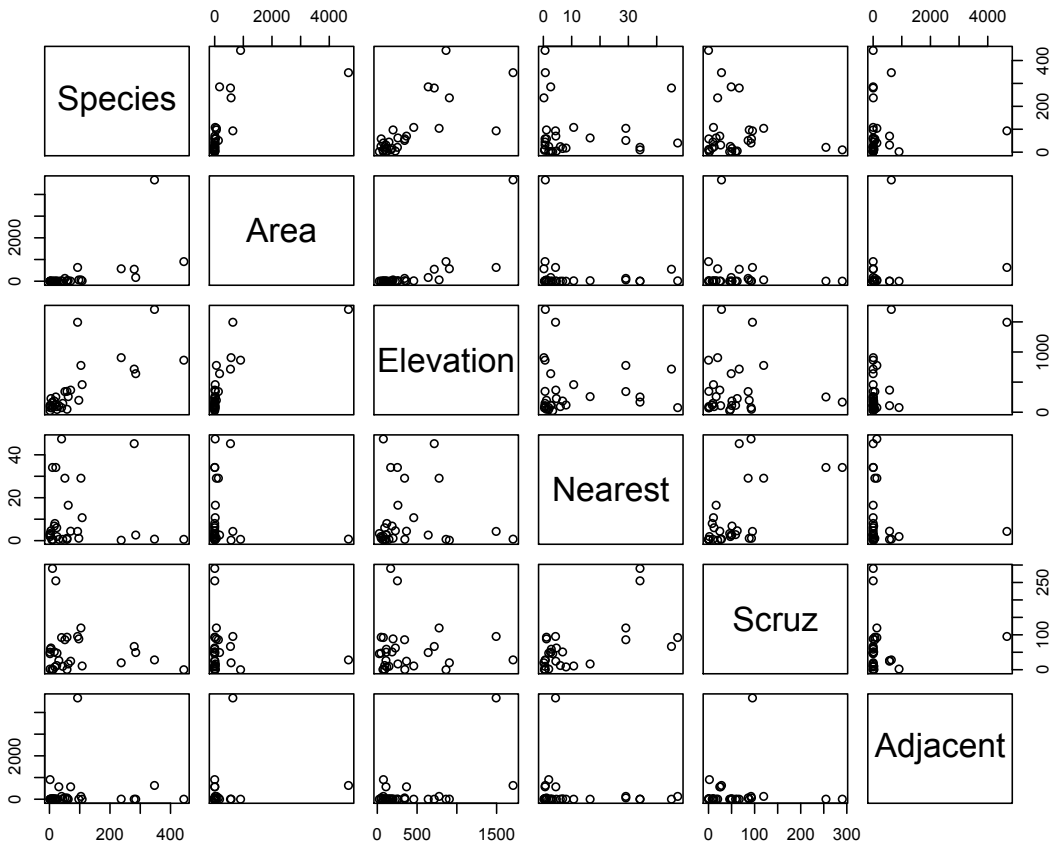
The column Species of the data set gala.txt indicates the number of different plant species on the Galapagos island. The explanatory variables are Area (area of island), Elevation (highest elevation of island), Nearest (distance to nearest island), Scruz (distance to Santa Cruz) and Adjacent (area of adjacent island). All explanatory variables are numeric.

features

y = resp.

```
> gala=read.table("gala.txt",header=TRUE); gala
             Species  Area Elevation Nearest Scruz Adjacent
Baltra            58 25.09       346     0.6   0.6     1.84
Bartolome         31  1.24       109     0.6  26.3   572.33
Caldwell           3  0.21       114     2.8  58.7     0.78
Champion          25  0.10        46     1.9  47.4     0.18
Coamano            2  0.05        77     1.9   1.9   903.82
Daphne.Major      18  0.34       119     8.0   8.0     1.84
[ some output deleted ]
```

Analysis in R: graphics



The problem of collinearity amongst explanatory variables is similar in nature as in the linear models case.

Analysis in R: estimation and testing

```
> galaglm=glm(Species~Area+Elevation+Nearest+Scruz+Adjacent,
+ family=poisson,data=gala)
> summary(galaglm)
[some output deleted ]
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area       -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation   3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest     8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz      -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent   -6.630e-04  2.933e-05 -22.608  < 2e-16 ***
```

y

predictors

est.s of coeff.s

p-val.s to test if corresp. β is equal to 0

μ

The output of the function `glm` is an object of type`glm`, to which functions as `anova`, `drop1`, `summary`, `coef`, `fitted`, `predict`, `confint`, etc. can be applied, in the same way as for the logistic regression. Remember that the interpretation of the predicted responses is of course different: for example, the predicted responses (i.e., estimate of EY_{in}) or the Poisson regression with one factor (with I levels) and one contin. predictor X are $\hat{Y}_{in} = \hat{\lambda}_{in} = e^{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}X_{in}}$.

predict command

for logistic regr. you get prob.s of succ.

further designs

Further designs

- Other GLM's for non-normal outcomes. Besides binomial and count data the `glm` function can also model multinomial, negative binomial, Gamma.
- Longitudinal analysis. In longitudinal experiments one is interested in the development of individuals or other experimental units over time. This typically leads to multiple measurements per individual, taken at different time points (and often modeled with mixed effects models).

used in medical studies

- Mixed models. Mixed models define outcomes in terms of parameters, (random) errors and additional random effects. This allows to model variation due to the selection of experimental units, fluctuations over time, extraneous variables that influence some measurements, etc.

To finish

Today we discussed

1. generalized linear models
    1. logistic regression
    2. Poisson regression