Part I - Generalized Linear Models:

An Introduction based on R

Herwig Friedl

Institute of Statistics
Graz University of Technology, Austria

hfriedl@tugraz.at

http:
//www.stat.tugraz.at/courses/glmLjubljana.html

May 2021

Introduction

- This course will provide an introduction into the concepts of the class of generalized linear models (GLM's).
- This class extends the class of linear models (LM's) to regression models for non-normal data.
- Special interest will be on binary data (logistic regression) and count data (log-linear models).
- All models will be handled by using R functions like `lm`, `anova`, or `glm`.

Plan

- Linear Models (LM's): Recap of Results
- Box-Cox Transformation Family: Extending the LM
- Generalized Linear Models (GLM's): An Introduction
- Linear Exponential Family (LEF): Properties and Members
- GLM's: Parameter Estimates
- GLM's: `glm(.)` Function
- Gamma Models
- Logistic Models (Binomial Frequencies)
- Log-linear Models (Poisson Counts)
- Multilevel Models

Recap Linear Models

Goal of regression models is to find out how a **response variable** depends on **covariates** (explanatory variables).
A special class of regression models are linear models. The general setup is given by

- Data $(y_i, x_{i1}, \ldots, x_{i,p-1})$, $i = 1, \ldots, n$
- Response $\mathbf{y} = (y_1, \ldots, y_n)^\top$ (random variable)
- Covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{i,p-1})^\top$ (fixed, known)

Recap Linear Models

Data Example: Life Expectancies

Data source: The World Bank makes available data from the **W**orld **D**evelopment **I**ndicators. To search/download within R:

```
> install.packages(’WDI’); library(WDI)
> WDIsearch(’gdp’) # gives a list of available data on gdp

> d <- WDI(indicator=’NY.GDP.PCAP.KD’, country=c(’AT’, ’US’),
+          start=1960, end=2013)
> head(d)
  iso2c country NY.GDP.PCAP.KD year
1    AT Austria       47901.37 2013
2    AT Austria       48172.24 2012
3    AT Austria       48065.32 2011
4    AT Austria       46858.04 2010
5    AT Austria       46123.49 2009
6    AT Austria       48053.48 2008
```

Recap Linear Models

Data Example: Life Expectancies

Data on `temperature` are available at *The World Bank, Climate Change Knowledge Portal: Historical Data*

```
> install.packages('gdata')
> library(gdata)
> f.name<-"http://databank.worldbank.org/data/download/catalog/
+          cckp_historical_data_0.xls"
> myperl <- "c:/Strawberry/perl/bin/perl.exe"
> sheetCount(f.name, perl=myperl)
Downloading...
trying URL 'http://databank.worldbank.org/data/.../*.xls'
Content type 'application/vnd.ms-excel' length 378368 bytes
opened URL
downloaded 369 Kb
Done.
[1] 5
```

Recap Linear Models

Data Example: Life Expectancies

```
> temp <- read.xls(f.name, sheet="Country_temperatureCRU",
+                  perl=myperl)
> temp.data <- temp[ , c("ISO_3DIGIT", "Annual_temp")]
> colnames(temp.data) <- c("iso3c", "temp")
> head(temp.data)
  iso3c       temp
1   AFG       12.92
2   AGO       21.51
3   ALB       11.27
4   ARE       26.83
5   ARG       14.22
6   ARM        6.37
```

Recap Linear Models

Data Example: Life Expectancies

Data we are interested in (from 2010):

- `life.exp` at birth, total (years)
- `urban` population (percent)
- `physicians` (per 1,000 people)
- `temp` annual mean (Celsius)

Which is the response and which are covariates?

Recap Linear Models

Gaussian Linear Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i \, , \quad \epsilon_i \overset{iid}{\sim} \mathrm{Normal}(0, \sigma^2) \, ,$$

with unknown **regression parameters** $\beta_0, \beta_1, \ldots, \beta_{p-1}$ (intercept β_0, slopes β_j, $j = 1, \ldots, p-1$) and unknown (homogenous) **error variance** σ^2.

This is equivalent with $y_i \overset{ind}{\sim} \mathrm{Normal}(\mathrm{E}(y_i), \mathrm{var}(y_i))$, where

$$\mathrm{E}(y_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}$$

is a **linear** function in the parameters and

$$\mathrm{var}(y_i) = \sigma^2 \, , \qquad i = 1, \ldots, n$$

describes a **homoscedastic** scenario.

Recap Linear Models

Matrix Notation: we define

$$\begin{aligned}
&\mathbf{y} = (y_1, \ldots, y_n)^\top, \quad \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top, \\
&\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top, \quad \mathbf{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})^\top, \\
&\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top
\end{aligned}$$

and write a Gaussian regression models as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with

$$\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

and

$$\mathrm{var}(\mathbf{y}) = \sigma^2 \mathbf{I}_n .$$

Here $\mathbf{I}_n$ denotes the $(n \times n)$ identity matrix, and the $(n \times p)$ matrix $\mathbf{X}$ is also called **Design Matrix**.

Recap Linear Models

Exploratory Data Analysis (EDA):

- Check out the **ranges** of the response and covariates. For **discrete** covariates (with sparse factor levels) we consider **grouping** the levels.

- Plot covariates against response. Scatter plot should reflect **linear** relationships otherwise we consider **transformations**.

- To check if the constant variance assumption is reasonable, the points of the scatter plot of covariates against the responses should be contained in a **band of constant width**.

Recap Linear Models

Data Example: Life Expectancies (EDA)

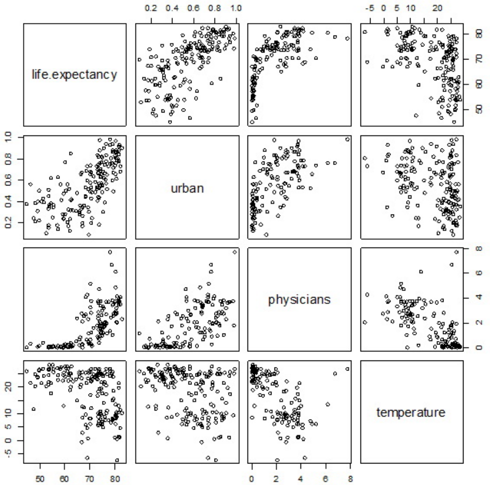
```
> summary(mydata[, c(5, 6, 8, 10)])
```

```
life.expectancy     urban            physicians       temperature
Min.   :45.10   Min.   :0.1064   Min.   :0.0080   Min.   :-7.14
1st Qu.:62.19   1st Qu.:0.3890   1st Qu.:0.2318   1st Qu.:10.40
Median :72.04   Median :0.5683   Median :1.4567   Median :21.90
Mean   :69.48   Mean   :0.5648   Mean   :1.6678   Mean   :18.24
3rd Qu.:76.03   3rd Qu.:0.7496   3rd Qu.:2.8146   3rd Qu.:25.06
Max.   :82.84   Max.   :1.0000   Max.   :6.8152   Max.   :28.30
                                 NA's   :23
```

Recap Linear Models

Data Example: Life Expectancies (EDA)

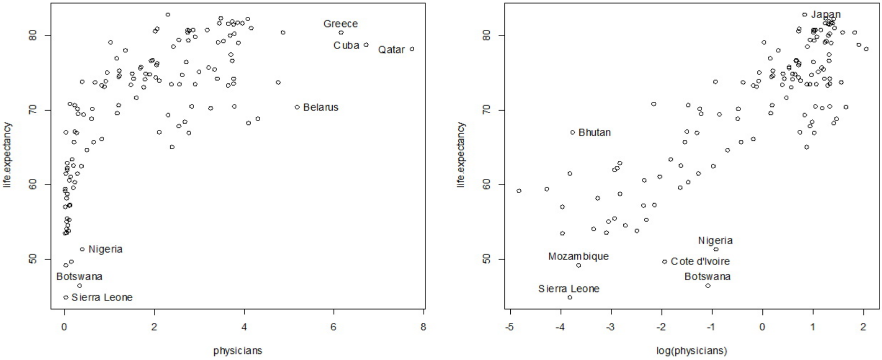
```
> plot(mydata[, c(5, 6, 8, 10)])
```



Recap Linear Models

Data Example: Life Expectancies (Transformations)

```
plot(physicians, life.expectancy)
plot(log(physicians), life.expectancy)
```



Recap Linear Models

Parameter Estimation: $\boldsymbol{\beta}$

Idea of **Least Squares**: minimize the sum of squared errors, i.e.

$$\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$$

Equivalent with **Maximum Likelihood**: maximize the sample log-likelihood function

$$\ell(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n} \left(\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 \right)$$

LSE/MLE **Solution**: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$

For $y_i \overset{ind}{\sim} \mathrm{Normal}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$ we have

$$\hat{\boldsymbol{\beta}} \sim \mathrm{Normal}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$$

Recap Linear Models

Parameter Estimation: $\boldsymbol{\sigma}^2$
Maximum Likelihood Estimator:

$$\hat{\sigma}^2 = \frac{1}{n}\,\mathrm{SSE}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}})^2, \qquad \mathrm{E}(\hat{\sigma}^2) = \left(1 - \frac{p}{n}\right)\sigma^2$$

is biased. An **unbiased** variance estimator is (df corrected)

$$S^2 = \frac{1}{n-p}\,\mathrm{SSE}(\hat{\boldsymbol{\beta}})$$

For $y_i \overset{ind}{\sim} \mathrm{Normal}(\mathbf{x}_i^\top\boldsymbol{\beta}, \sigma^2)$ we get

$$\mathrm{SSE}(\hat{\boldsymbol{\beta}})/\sigma^2 \sim \chi^2_{n-p}$$

and $\mathrm{SSE}(\hat{\boldsymbol{\beta}})$ is **stochastically independent** of $\hat{\boldsymbol{\beta}}$.

Recap Linear Models

ANalysis Of VAriance (ANOVA): let $\hat{\mu}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$, then

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2}_{\text{SSR}(\hat{\boldsymbol{\beta}})} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}_{\text{SSE}(\hat{\boldsymbol{\beta}})}$$

Total SS equals (maxim.) **Regression** SS plus (minim.) **Error** SS

Thus, the proportion of variability explained by the regression model is described by the **coefficient of determination**

$$R^2 = \frac{\text{SSR}(\hat{\boldsymbol{\beta}})}{\text{SST}} = 1 - \frac{\text{SSE}(\hat{\boldsymbol{\beta}})}{\text{SST}} \in (0, 1)$$

To penalize for model complexity p we use its **adjusted** version

$$R^2_{adj} = 1 - \frac{\text{SSE}(\hat{\boldsymbol{\beta}})/(n-p)}{\text{SST}\,/(n-1)} \notin (0, 1)$$

Recap Linear Models

Hypothesis Tests: t-Test

If the model is correctly stated then

$$\hat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$$

Thus, for each **slope** parameter β_j, $j = 1, \ldots, p-1$, we have

$$\hat{\beta}_j \sim \text{Normal}(\beta_j, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}_{j+1,j+1})$$

and therefore

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}_{j+1,j+1}}} \sim \text{Normal}(0,1)$$

Since S^2 and $\hat{\boldsymbol{\beta}}$ are independent, replacing σ^2 by S^2 results in

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{S^2(\mathbf{X}^\top\mathbf{X})^{-1}_{j+1,j+1}}} \sim t_{n-p}$$

Recap Linear Models

Hypothesis Tests: t-Test

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{S^2(\mathbf{X}^\top\mathbf{X})^{-1}_{j+1,j+1}}} \sim t_{n-p}$$

Therefore, we can test the relevance of a **single predictor** x_j by

$$H_0 : \beta_j = 0 \qquad \text{vs} \qquad H_1 : \beta_j \neq 0$$

and use the well-known **test statistic**

$$\frac{\text{Estimate}}{\text{Std. Error}} = \frac{\hat{\beta}_j}{\sqrt{S^2(\mathbf{X}^\top\mathbf{X})^{-1}_{j+1,j+1}}} \overset{H_0}{\sim} t_{n-p}$$

Recap Linear Models

Hypothesis Tests: F-Test

If a predictor is a **factor** with k levels (e.g., `continent`: Europe, Africa, America, Asia), then we usually define a baseline category (e.g. Europe) and consider the model

$$\mu = \beta_0 + \beta_{Af} I(Africa) + \beta_{Am} I(America) + \beta_{As} I(Asia)$$

To check if the predictor `continent` is irrelevant we have to **simultaneously** test $k-1$ parameters

$$H_0 : \beta_{Af} = \beta_{Am} = \beta_{As} = 0 \qquad \text{vs} \qquad H_1 : not\ H_0$$

Fitting the model twice, under H_0 and under H_1, results in $\mathrm{SSR}(\hat{\boldsymbol{\beta}}_0)$ and $\mathrm{SSR}(\hat{\boldsymbol{\beta}}_1)$ and we get the **test statistic**

$$\frac{\left(\mathrm{SSR}(\hat{\boldsymbol{\beta}}_1) - \mathrm{SSR}(\hat{\boldsymbol{\beta}}_0)\right)/(k-1)}{\mathrm{SSE}(\hat{\boldsymbol{\beta}}_1)/(n-p)} \overset{H_0}{\sim} F_{k-1,n-p}.$$

Recap Linear Models

Weighted Least Squares in case of heteroscedastic errors, i.e.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, \sigma^2 \mathbf{W}), \quad \mathbf{W} = \text{diag}(w_1, \ldots, w_n)$$

The MLE (weighted LSE) of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{-1} \mathbf{y}$$

with

$$\text{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \quad \text{and} \quad \text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{W}^{-1} \mathbf{X})^{-1}$$

The MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{w_i} = \frac{1}{n} \mathbf{r}^\top \mathbf{W}^{-1} \mathbf{r}$$

with the vector of raw residuals $\mathbf{r} = \mathbf{y} - \hat{\boldsymbol{\mu}}$.

Recap Linear Models

Data Example: Life Expectancies

```
> mod <- lm(life.expectancy ~ urban + physicians + temperature)
> summary(mod)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.61188    2.01497  29.088  < 2e-16 ***
urban       14.66519    2.72913   5.374 3.09e-07 ***
physicians   2.72412    0.50569   5.387 2.90e-07 ***
temperature -0.07181    0.06758  -1.063     0.29
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.459 on 142 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.6191, Adjusted R-squared:  0.611
F-statistic: 76.93 on 3 and 142 DF,  p-value: < 2.2e-16
```

Recap Linear Models

Data Example: Life Expectancies

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 58.61188    2.01497  29.088  < 2e-16 ***
urban       14.66519    2.72913   5.374 3.09e-07 ***
physicians   2.72412    0.50569   5.387 2.90e-07 ***
temperature -0.07181    0.06758  -1.063     0.29    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The predictors `urban` and `physicians` are significant. Only `temperature` has a negative effect and is also not significant.

Recap Linear Models

Data Example: Life Expectancies

```
Residual standard error: 5.459 on 142 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.6191, Adjusted R-squared:  0.611
F-statistic: 76.93 on 3 and 142 DF,  p-value: < 2.2e-16
```

Under the model, the estimated standard error of the response is 5.5 (years). We have $n - p = 142$ and $p - 1 = 3$ predictors.

Almost 62% of the total variability is explained by this model. The adjusted version of R^2 is 61.1%.

We finally test that **all three** predictors are irrelevant. The associated F-test clearly rejects this hypothesis.

Recap Linear Models

Data Example: Life Expectancies (log(physicians))

```
> mod.log <- update(mod, .~. -physicians+log(physicians))
> summary(mod.log)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     66.70367    1.79065  37.251  < 2e-16 ***
urban            8.76445    2.53243   3.461 0.000711 ***
temperature     -0.03008    0.05668  -0.531 0.596408
log(physicians)  3.51370    0.39341   8.931 1.97e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Predictor `log(physicians)` is now highly significant but `temperature` lost it's significance!

Recap Linear Models

Data Example: Life Expectancies (log(physicians))

```
Residual standard error: 4.794 on 142 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.7063, Adjusted R-squared:  0.7001
F-statistic: 113.8 on 3 and 142 DF,  p-value: < 2.2e-16
```

Standard error is much smaller now than before (±4.8 years)!

Even 70% of the total variability is now explained by this model.

Same conclusion based on global F-test as in previous model.

Recap Linear Models

Data Example: Life Expectancies (ANOVA)

```
> anova(mod.log)
Analysis of Variance Table

Response: life.expectancy
                 Df Sum Sq Mean Sq F value    Pr(>F)
urban             1 5359.7  5359.7 233.219 < 2.2e-16 ***
temperature       1  653.2   653.2  28.424 3.747e-07 ***
log(physicians)   1 1833.3  1833.3  79.771 1.973e-15 ***
Residuals       142 3263.4    23.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Recap Linear Models

ANOVA

Remember the SST decomposition under the **Model** $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$:

$$\mathrm{SST} = \mathrm{SSR}(\hat{\boldsymbol{\beta}}) + \mathrm{SSE}(\hat{\boldsymbol{\beta}})$$

Information about this is contained in the **ANOVA Table**:

Source	df	Sum of Sq.	MSS	F
Regression	$p-1$	$\mathrm{SSR}(\hat{\boldsymbol{\beta}})$	$\mathrm{MSR}(\hat{\boldsymbol{\beta}}) = \mathrm{SSR}(\hat{\boldsymbol{\beta}})/(p-1)$	$\dfrac{\mathrm{MSR}(\hat{\boldsymbol{\beta}})}{\mathrm{MSE}(\hat{\boldsymbol{\beta}})}$
Error	$n-p$	$\mathrm{SSE}(\hat{\boldsymbol{\beta}})$	$\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \mathrm{SSE}(\hat{\boldsymbol{\beta}})/(n-p)$	
Total	$n-1$	SST		

Recap Linear Models

ANOVA

Null Model: assuming an **iid** random sample ($\mathrm{E}(y_i) = \beta_0$), results in $\mathrm{SSE}(\hat{\beta}_0) = \sum_i (y_i - \hat{\beta}_0)^2$ with $\hat{\beta}_0 = \bar{y}$. Thus, $\mathrm{SSE}(\hat{\beta}_0) = \sum_i (y_i - \bar{y})^2 \equiv \mathrm{SST}$ in this case.

Nested Model: we assume that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \quad \text{and test on} \quad H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$$

with $\dim(\boldsymbol{\beta}_1) = p_1$ (including the intercept) and $\dim(\boldsymbol{\beta}_2) = p_2$ (additional slopes). The corresponding SSR and SSE terms are

$$\mathrm{SSR}(\hat{\boldsymbol{\beta}}_1) = \sum_{i=1}^{n} (\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_1 - \bar{y})^2, \qquad \mathrm{SSE}(\hat{\boldsymbol{\beta}}_1) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_1)^2$$

Recap Linear Models

ANOVA

Sequentially adding the term $\mathbf{X}_2$ in the model where $\mathbf{X}_1$ is already included results in

Source	df	Sum of Squares/SS	MSS	F
$\mathbf{X}_1$	$p_1 - 1$	$\text{SSR}(\hat{\boldsymbol{\beta}}_1)$	$\text{MSR}(\hat{\boldsymbol{\beta}}_1) = \frac{\text{SSR}(\hat{\boldsymbol{\beta}}_1)}{p_1 - 1}$	$\frac{\text{MSR}(\hat{\boldsymbol{\beta}}_1)}{\text{MSE}(\hat{\boldsymbol{\beta}})}$
$\mathbf{X}_2 \vert \mathbf{X}_1$	p_2	$\text{SSR}(\hat{\boldsymbol{\beta}}_2 \vert \hat{\boldsymbol{\beta}}_1) =$ $\text{SSR}(\hat{\boldsymbol{\beta}}) - \text{SSR}(\hat{\boldsymbol{\beta}}_1)$	$\text{MSR}(\hat{\boldsymbol{\beta}}_2 \vert \hat{\boldsymbol{\beta}}_1) = \frac{\text{SSR}(\hat{\boldsymbol{\beta}}_2 \vert \hat{\boldsymbol{\beta}}_1)}{p_2}$	$\frac{\text{MSR}(\hat{\boldsymbol{\beta}}_2 \vert \hat{\boldsymbol{\beta}}_1)}{\text{MSE}(\hat{\boldsymbol{\beta}})}$
Error	$n - p$	$\text{SSE}(\hat{\boldsymbol{\beta}})$	$\text{MSE}(\hat{\boldsymbol{\beta}}) = \text{SSE}(\hat{\boldsymbol{\beta}})/(n - p)$	
Total	$n - 1$	SST		

Recap Linear Models

ANOVA

We now assume that the model $\mathbf{y} = \beta_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$ holds.

Test 1: test statistic

$$F = \frac{\text{MSR}(\hat{\boldsymbol{\beta}}_1|\hat{\beta}_0)}{\text{MSE}(\hat{\boldsymbol{\beta}})}$$

tests the model improvement when adding the predictors in $\mathbf{X}_1$ to the iid model based on β_0 only.

Test 2: test statistic

$$F = \frac{\text{MSR}(\hat{\boldsymbol{\beta}}_2|\hat{\boldsymbol{\beta}}_1, \hat{\beta}_0)}{\text{MSE}(\hat{\boldsymbol{\beta}})}$$

tests the model improvement when adding the predictors in $\mathbf{X}_2$ to the model with $\mathbf{X}_1$ and β_0 already contained.

Recap Linear Models

Data Example: Life Expectancies (ANOVA)

```
> anova(mod.log)
Analysis of Variance Table

Response: life.expectancy
                 Df Sum Sq Mean Sq F value    Pr(>F)
urban             1 5359.7  5359.7 233.219 < 2.2e-16 ***
temperature       1  653.2   653.2  28.424 3.747e-07 ***
log(physicians)   1 1833.3  1833.3  79.771 1.973e-15 ***
Residuals       142 3263.4    23.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Each further predictor that enters the model significantly improves the model fit.

Linear Models: Restrictions

Problems:

- $y_i \not\sim \text{Normal}(\text{E}(y_i), \text{var}(y_i))$
- $\text{E}(y_i) \neq \mathbf{x}_i^\top \boldsymbol{\beta} \in \mathbb{R}$
- $\text{var}(y_i) \neq \sigma^2$ equal (homoscedastic) for all $i = 1, \ldots, n$

Remedies:

- transform y_i such that $g(y_i) \overset{ind}{\sim} \text{Normal}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$
- utilize a GLM where $y_i \overset{ind}{\sim} \text{LEF}(g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}), \phi V(\mu_i))$

Box-Cox Transformation

For **positive** Responses ($y > 0$) define

$$y(\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log y, & \text{if } \lambda = 0, \end{cases}$$

$y(\lambda) \to \log y$ for $\lambda \to 0$, such that $y(\lambda)$ is continuous in λ.

Assumption: there is a value λ for which

$$y_i(\lambda) \overset{ind}{\sim} \text{Normal}\Big(\mu_i(\lambda) = \mathbf{x}_i^\top \boldsymbol{\beta}(\lambda), \sigma^2(\lambda)\Big)$$

Compute **MLEs** with respect to the sample density of the **untransformed** (original) response y.

Box-Cox Transformation

Density Transformation Theorem: If $g(Y) \sim \mathrm{F}_{g(Y)}(y)$ holds for a continuous r.v. and $g(\cdot)$ is a monotone function, then the untransformed r.v. Y has cdf

$$F_Y(y) = \Pr(Y \leq y) = \Pr(g(Y) \leq g(y)) = F_{g(Y)}(g(y)).$$

Thus, the density of Y is

$$f_Y(y) = \frac{\partial F_{g(Y)}(g(y))}{\partial y} = f_{g(Y)}(g(y)) \cdot \left| \frac{\partial g(y)}{\partial y} \right|$$

with Jacobian $\left| \frac{\partial g(y)}{\partial y} \right|$.

Box-Cox Transformation

Density of untransformed y is

$$f(y|\lambda,\mu(\lambda),\sigma^2(\lambda)) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2(\lambda)}} \exp\left(-\frac{\left(\frac{(y^\lambda-1)}{\lambda}-\mu(\lambda)\right)^2}{2\sigma^2(\lambda)}\right) y^{\lambda-1}, & \lambda \neq 0, \\ \frac{1}{\sqrt{2\pi\sigma^2(\lambda)}} \exp\left(-\frac{(\log y-\mu(\lambda))^2}{2\sigma^2(\lambda)}\right) y^{-1}, & \lambda = 0. \end{cases}$$

- If $\lambda \neq 0$ and $\mu(\lambda) = \mathbf{x}^\top \boldsymbol{\beta}(\lambda)$ then

$$f(y|\lambda,\mu(\lambda),\sigma^2(\lambda)) = \frac{1}{\sqrt{2\pi\lambda^2\sigma^2(\lambda)}} \exp\left(-\frac{\left(y^\lambda - 1 - \lambda\mathbf{x}^\top\boldsymbol{\beta}(\lambda)\right)^2}{2\lambda^2\sigma^2(\lambda)}\right) |\lambda| y^{\lambda-1}.$$

Box-Cox Transformation

Using $\beta_0 = 1 + \lambda\beta_0(\lambda)$, $\beta_j = \lambda\beta_j(\lambda)$, $j = 1, \dots, p-1$, and $\sigma^2 = \lambda^2\sigma^2(\lambda)$ then

$$f(y|\lambda, \mu(\lambda), \sigma^2(\lambda)) = \frac{1}{\sqrt{2\pi\lambda^2\sigma^2(\lambda)}} \exp\left(-\frac{\left(y^\lambda - 1 - \lambda\mathbf{x}^\top\boldsymbol{\beta}(\lambda)\right)^2}{2\lambda^2\sigma^2(\lambda)}\right)|\lambda|y^{\lambda-1}$$

$$f(y|\lambda, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^\lambda - \mathbf{x}^\top\boldsymbol{\beta})^2}{2\sigma^2}\right)|\lambda|y^{\lambda-1}.$$

- If $\lambda = 0$, let $\beta_j = \beta_j(\lambda)$, $j = 0, \dots, p-1$, and $\sigma^2 = \sigma^2(\lambda)$

$$f(y|0, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log y - \mathbf{x}^\top\boldsymbol{\beta})^2}{2\sigma^2}\right) y^{-1}.$$

If λ would be known, then the MLE could be easily computed!

Box-Cox Transformation

Relevant part of the sample log-likelihood function is

- $\lambda \neq 0$:

$$\ell(\lambda, \boldsymbol{\beta}, \sigma^2|\mathbf{y}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i^{\lambda} - \mathbf{x}_i^{\top}\boldsymbol{\beta}\right)^2 + n\log|\lambda| + (\lambda-1)\sum_{i=1}^{n}\log y_i$$

- $\lambda = 0$:

$$\ell(0, \boldsymbol{\beta}, \sigma^2|\mathbf{y}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\log y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta}\right)^2 - \sum_{i=1}^{n}\log y_i$$

Box-Cox Transformation: MLE's

If λ would be known, then the MLEs would be

$$\hat{\boldsymbol{\beta}}_\lambda = \begin{cases} (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}^\lambda, & \lambda \neq 0, \\ (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \log \mathbf{y}, & \lambda = 0, \end{cases}$$

$$\hat{\sigma}^2_\lambda = \frac{1}{n}\mathrm{SSE}_\lambda(\hat{\boldsymbol{\beta}}_\lambda) = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(y_i^\lambda - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_\lambda)^2, & \lambda \neq 0, \\ \frac{1}{n}\sum_{i=1}^{n}(\log y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_\lambda)^2, & \lambda = 0. \end{cases}$$

Box-Cox Transformation: Profile-Likelihood

Profile (log-) likelihood function $p\ell(\lambda|\mathbf{y}) = \ell(\lambda, \hat{\boldsymbol{\beta}}_\lambda, \hat{\sigma}^2_\lambda|\mathbf{y}) =$

$$= \begin{cases} -\frac{n}{2}\log \mathrm{SSE}_\lambda(\hat{\boldsymbol{\beta}}_\lambda) + n\log|\lambda| + (\lambda-1)\sum_{i=1}^{n}\log y_i, & \lambda \neq 0, \\ -\frac{n}{2}\log \mathrm{SSE}_0(\hat{\boldsymbol{\beta}}_0) - \sum_{i=1}^{n}\log y_i, & \lambda = 0. \end{cases}$$

This is the sample log-likelihood function that has been already maximized with respect to β and σ^2.

It only depends on the transformation parameter λ.

Find the maximum in λ by simply using a grid search strategy.

Box-Cox Transformation: Profile-Likelihood

Likelihood Ratio Test (LRT): $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$. For the LRT statistic it holds that

$$-2\left(p\ell(\lambda_0|\mathbf{y}) - p\ell(\hat{\lambda}|\mathbf{y})\right) \overset{D}{\rightarrow} \chi^2_1.$$

If $-2(p\ell(\lambda_0|\mathbf{y}) - p\ell(\hat{\lambda}|\mathbf{y})) \sim \chi^2_1$, a $(1-\alpha)$ confidence interval contains all values λ_0, for which

$$-(p\ell(\lambda_0|\mathbf{y}) - p\ell(\hat{\lambda}|\mathbf{y})) < \frac{1}{2}\chi^2_{1;1-\alpha}$$

(notice that $\chi^2_{1;0.95} = 3.841$, $\chi^2_{1;0.99} = 6.635$).

Box-Cox Transformation: Properties

Log-Transformation ($\lambda = 0$): if $\log y_i \sim \text{Normal}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$ then

$$\begin{aligned} \text{median}(\log y_i) &= \mathbf{x}_i^\top \boldsymbol{\beta}, \\ \text{E}(\log y_i) &= \mathbf{x}_i^\top \boldsymbol{\beta}, \\ \text{var}(\log y_i) &= \sigma^2. \end{aligned}$$

Untransformed response y_i follows a log-normal distribution with

$$\begin{aligned} \text{median}(y_i) &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta}), \\ \text{E}(y_i) &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta} + \sigma^2/2) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \exp(\sigma^2/2), \\ \text{var}(y_i) &= \big(\exp(\sigma^2) - 1\big) \exp(2\mathbf{x}_i^\top \boldsymbol{\beta} + \sigma^2). \end{aligned}$$

- **Additive** model for mean and median of $\log y_i$ corresponds to a **multiplicative** model for mean and median of y_i.
- $\text{E}(y_i)$ is $1 < \exp(\sigma^2/2)$ times its $\text{median}(y_i)$.
- $\text{var}(y_i)$ is no longer constant for $i = 1, \ldots, n$.

Box-Cox Transformation: Properties

Power-Transformation ($\lambda \neq 0$): if $y_i^\lambda \sim \text{Normal}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$ then

$$\begin{aligned} \text{median}(y_i^\lambda) &= \mathbf{x}_i^\top \boldsymbol{\beta}, \\ \text{E}(y_i^\lambda) &= \mathbf{x}_i^\top \boldsymbol{\beta}, \\ \text{var}(y_i^\lambda) &= \sigma^2. \end{aligned}$$

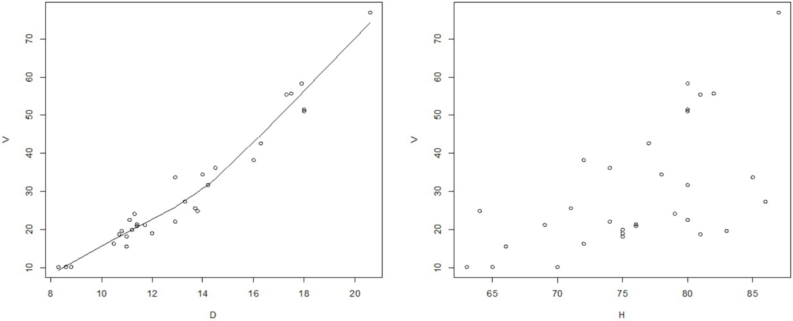
Untransformed response y_i follows a distribution with

$$\begin{aligned} \text{median}(y_i) &= \mu_i^{1/\lambda}, \\ \text{E}(y_i) &\approx \mu_i^{1/\lambda} \left(1 + \sigma^2(1-\lambda)/(2\lambda^2\mu_i^2)\right), \\ \text{var}(y_i) &\approx \mu_i^{2/\lambda}\sigma^2/(\lambda^2\mu_i^2). \end{aligned}$$

Box-Cox Transformation: Example

Girth (diameter), Height and Volume for $n = 31$ Black Cherry Trees available in R.
Relationship between volume V in feet^3, height H in feet and diameter D in inches (1 inch = 2.54 cm, 12 inches = 1 foot).

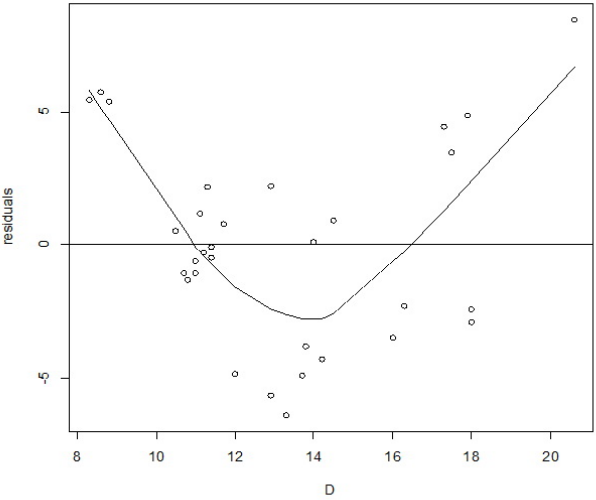
```
> H <- trees$Height; D <- trees$Girth; V <- trees$Volume
> plot(D, V); lines(lowess(D, V)) # curvature (wrong scale?)
> plot(H, V)  # increasing variance?
```



Box-Cox Transformation: Example

```
> (mod <- lm(V ~ H + D)) # still fit a linear model for volume
Coefficients:
(Intercept)            H            D
   -57.9877       0.3393       4.7082

> plot(D, residuals(mod), ylab="residuals"); abline(0, 0)
> lines(lowess(D, residuals(mod))) # sink in the middle
```

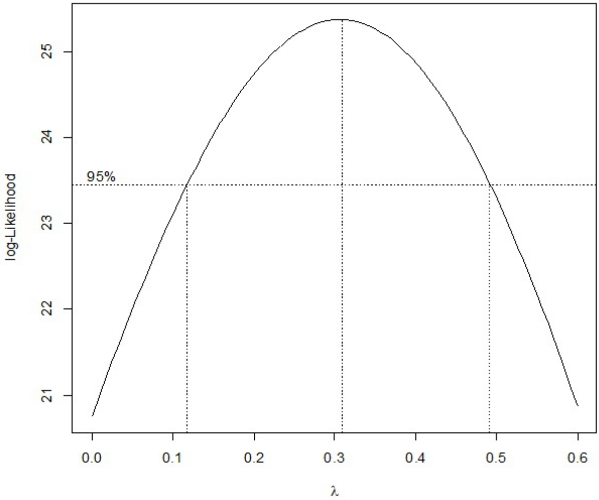
Box-Cox Transformation: Example



Box-Cox Transformation: Example

```
> library(MASS)

> bc<-boxcox(V~H+D,lambda=seq(0.0,0.6,length=100),plotit=FALSE)
> ml.index <- which(bc$y == max(bc$y))
> bc$x[ml.index]
[1] 0.3090909

> boxcox(V~H+D, lambda = seq(0.0, 0.6,len = 18)) # plot it now
```

Box-Cox Transformation: Example



Box-Cox Transformation: Example

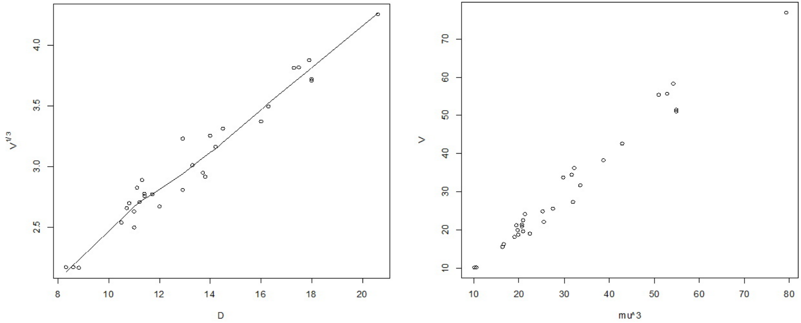
Is volume cubic in height and diameter?

```
> plot(D, V^(1/3), ylab=expression(V^{1/3}))
> lines(lowess(D, V^(1/3))) # curvature almost gone

> (mod1 <- lm(V^(1/3) ~ H + D))
Coefficients:
(Intercept)            H            D
   -0.08539      0.01447      0.15152
```

For fixed $\lambda = 1/3$ we have $\widehat{\text{median}}(V) = \hat{\mu}^3_{1/3}$ where $\text{E}(V^{1/3}) = \mu_{1/3}$. $\hat{\text{E}}(V) = \hat{\mu}^3_{1/3}(1 + 3\hat{\sigma}^2_{1/3}/\hat{\mu}^2_{1/3})$. Compare responses with estimated medians

```
> mu <- fitted(mod1)
> plot(mu^3, V) # fitted median modell
```

Box-Cox Transformation: Example

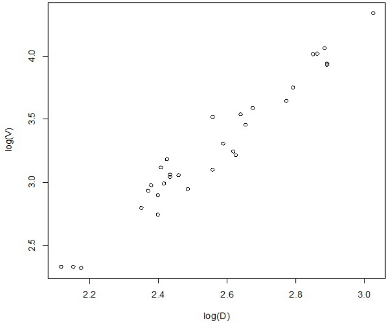


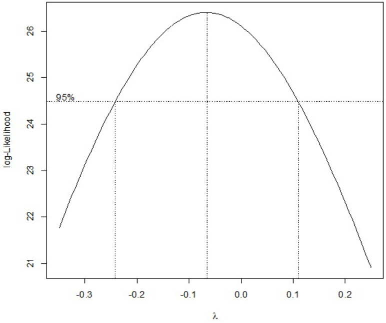
Box-Cox Transformation: Example

Alternative strategy:
Remove curvature by a log-transform of all predictors (i.e., regress on $\log(D)$ and $\log(H)$).
Should we also consider $\log(V)$ as response?

```
> plot(log(D), log(V)) # shows nice linear relationship

> lm(log(V) ~ log(H) + log(D)) # response log(V) or still V?
Coefficients:
(Intercept)       log(H)       log(D)
     -6.632        1.117        1.983

> boxcox(V~log(H)+log(D), lambda=seq(-0.35,0.25,length=100))
```

Box-Cox Transformation: Example





Box-Cox Transformation: Example

Which of the models is *better*? Comparison by LRT. Both models are members of the **model family**

$$V^* \sim \text{Normal}(\beta_0 + \beta_1 H^* + \beta_2 D^*, \sigma^2)$$

$$\begin{aligned} V^* &= (V^{\lambda_V} - 1)/\lambda_V \\ H^* &= (H^{\lambda_H} - 1)/\lambda_H \\ D^* &= (D^{\lambda_D} - 1)/\lambda_D \end{aligned}$$

Compare Profile-Likelihood function in $\lambda_V = 1/3$, $\lambda_H = \lambda_D = 1$ $(E(V^{1/3}) = \beta_0 + \beta_1 H + \beta_2 D)$, with that in $\lambda_V = \lambda_H = \lambda_D = 0$ $(E(\log(V)) = \beta_0 + \beta_1 \log(H) + \beta_2 \log(D))$.

Box-Cox Transformation: Example

```
> bc1 <- boxcox(V ~ H + D, lambda = 1/3, plotit=FALSE)
> bc1$y
[1] 25.33313

> bc2 <- boxcox(V ~ log(H) + log(D), lambda = 0, plotit=FALSE)
> bc2$y
[1] 26.11592
```

LRT Statistic: $-2(25.333 - 26.116) = 1.566$ (**not significant**).

Box-Cox Transformation: Example

Remark: Coefficient of $\log(H)$ close to 1 ($\hat{\beta}_1 = 1.117$) and coefficient of $\log(D)$ close to 2 ($\hat{\beta}_2 = 1.983$).

Tree can be represented by a **cylinder** or a **cone**. Volume is $\pi hd^2/4$ (cylinder) or $\pi hd^2/12$ (cone), i.e.

$$\mathrm{E}(\log(V)) = c + 1\log(H) + 2\log(D)$$

with $c = \log(\pi/4)$ (cylinder) or $c = \log(\pi/12)$ (cone).

Attention: D has to be converted from inches to feet $\Rightarrow D/12$ as predictor.

Box-Cox Transformation: Example

```
> lm(log(V) ~ log(H) + log(D/12))
Coefficients:
(Intercept)       log(H)    log(D/12)
     -1.705        1.117        1.983
```

Conversion only influences intercept!

Fix slopes (β_1, β_2) to (1, 2) and estimate only intercept β_0, i.e. consider the model

$$E(\log(V)) = \beta_0 + 1\log(H) + 2\log(D/12)\,.$$

Term $1\log H + 2\log(D/12)$ is called **offset** (predictor with fixed parameter 1).

Box-Cox Transformation: Example

```
> (mod3 <- lm(log(V) ~ 1 + offset(log(H) + 2*log(D/12))))
Coefficients:
(Intercept)
     -1.199

> log(pi/4)
 [1] -0.2415645
> log(pi/12)
 [1] -1.340177
```

Volume can be better described by a cone than by a cylinder. However, its volume is slightly larger than the one of a cone.

Introduction to GLM's

- In **generalized linear models** (GLM's) we again have independent response variables with covariates.
- While a linear model combines **additivity** of the covariate effects with the **normality** of the errors, including **variance homogeneity**, GLM's don't need to satisfy these requirements. GLM's allow also to handle **nonnormal** responses such as binomial, Poisson and Gamma.
- Regression parameters are estimated using **maximum likelihood**.
- Standard reference on GLM's is McCullagh & Nelder (1989).

Introduction to GLM's: Components of a GLM

Response y_i and covariables $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})^\top$.

1. Random Component:
   y_i, $i = 1, \ldots, n$, independent with density from the **linear exponential family (LEF)**, i.e.

   $$f(y|\theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\}$$

   $\phi > 0$ is a dispersion parameter and $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions.
2. Systematic Component:
   $\eta_i = \eta_i(\boldsymbol{\beta}) = \mathbf{x}_i^\top \boldsymbol{\beta}$ is called **linear predictor**,
   $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\top$ are unknown regression parameters
3. Parametric Link Component:
   The **link function** $g(\mu_i) = \eta_i$ combines the linear predictor with the mean of y_i. **Canonical** link function if $\theta = \eta$.

Introduction to GLM's: LM as GLM

$y_i \sim \text{Normal}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$, independent, $i = 1, \ldots, n$. Density has LEF form, since

$$
\begin{aligned}
f(y|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \\
&= \exp\left\{\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left[\log(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\right]\right\}
\end{aligned}
$$

Defining $\theta = \mu$ and $\phi = \sigma^2$ results in

$$
b(\theta) = \frac{\mu^2}{2} \quad \text{and} \quad c(y, \phi) = -\frac{1}{2}\left[\log(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\right]
$$

Since $\theta = \mu$, the canonical link $g(\mu) = \mu$ is used in a LM.

Introduction to GLM's: Moments

It can be shown that for the LEF

$$\begin{aligned} \mathrm{E}(y) &= b'(\theta) = \mu \\ \mathrm{var}(y) &= \phi b''(\theta) = \phi V(\mu), \end{aligned}$$

where $V(\mu) = b''(\theta)$ is called the **variance function**.
Thus, we generally consider the model

$$g(\mu) = g(b'(\theta)).$$

Thus, the **canonical link** is defined as

$$\begin{aligned} g &= (b')^{-1} \\ \Rightarrow g(\mu) &= \theta = \mathbf{x}^{\top}\boldsymbol{\beta}. \end{aligned}$$

Introduction to GLM's: Estimating parameters

A single algorithm can be used to estimate the parameters of an LEM glm using **maximum likelihood**.
The log-likelihood of the sample $y_1, \ldots, y_n$ is

$$\ell(\boldsymbol{\mu}|\mathbf{y}) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}$$

The maximum likelihood estimator $\hat{\boldsymbol{\mu}}$ is obtained by solving the score function (chain rule)

$$s(\boldsymbol{\mu}) = \frac{\partial}{\partial \boldsymbol{\mu}} \ell(\boldsymbol{\mu}|\mathbf{y}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\mu}|\mathbf{y}) \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} = \left(\frac{y_1 - \mu_1}{\phi V(\mu_1)}, \ldots, \frac{y_n - \mu_n}{\phi V(\mu_n)} \right)$$

that only depends on a **mean/variance relationship**.

Introduction to GLM's: Estimating parameters

Because of $\mu = \mu(\boldsymbol{\beta})$ the score function for the parameter $\boldsymbol{\beta}$ is (chain rule again)

$$s(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}|\mathbf{y}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\mu}|\mathbf{y}) \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{1}{g'(\mu_i)} \mathbf{x}_i$$

which depends again only on the **mean/variance relationship**.

For the sample $y_1, \dots, y_n$ we assumed that there is only one **global dispersion parameter** ϕ, i.e. $\mathrm{E}(y_i) = \mu_i$, $\mathrm{var}(y_i) = \phi V(\mu_i)$.

Introduction to GLM's: Estimating parameters

The score equation to be solved for the MLE $\hat{\boldsymbol{\beta}}$ is

$$\sum_{i=1}^{n} \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)} \frac{1}{g'(\hat{\mu}_i)} \mathbf{x}_i = \mathbf{0}$$

which doesn't depend on ϕ and where $g(\hat{\mu}_i) = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$.
Notice, if a canonical link ($g(\mu) = \theta$) is used, we have

$$g'(\mu) = \frac{\partial \theta}{\partial \mu} = \frac{1}{\partial \mu / \partial \theta} = \frac{1}{\partial b'(\theta)/\partial \theta} = \frac{1}{b''(\theta)} = \frac{1}{V(\mu)}$$

and the above score equation simplifies to

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_i) \mathbf{x}_i = \mathbf{0}$$

Introduction to GLM's: Estimating parameters

A general method to solve the score equation is the iterative algorithm **Fisher's Method of Scoring** (derived from a Taylor expansion of $s(\boldsymbol{\beta})$).
In the t-th iteration, the new estimate $\boldsymbol{\beta}^{(t+1)}$ is obtained from the previous one $\boldsymbol{\beta}^{(t)}$ by

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + s(\boldsymbol{\beta}^{(t)}) \left[\mathrm{E}\left(\frac{\partial s(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} \right]^{-1}$$

Therefore, the speciality is the usage of the **expected** instead of the **observed** Hessian matrix.

Introduction to GLM's: Estimating parameters

It could be shown that this iteration can be rewritten as

$$\boldsymbol{\beta}^{(t+1)} = \left(\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{z}^{(t)}$$

with the vector of pseudo-observations $\mathbf{z} = (z_1, \dots, z_n)^\top$ and diagonal weight matrix $\mathbf{W}$ defined as

$$\begin{aligned} z_i &= g(\mu_i) + g'(\mu_i)(y_i - \mu_i) \\ w_i &= \frac{1}{V(\mu_i)(g'(\mu_i))^2} \end{aligned}$$

Introduction to GLM's: Estimating parameters

Since

$$\boldsymbol{\beta}^{(t+1)} = \left(\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{z}^{(t)}$$

the estimate $\hat{\boldsymbol{\beta}}$ is calculated using an **Iteratively (Re-)Weighted Least Squares** (IWLS) algorithm:

1. start with initial guesses $\mu_i^{(0)}$ (e.g. $\mu_i^{(0)} = y_i$ or $\mu_i^{(0)} = y_i + c$)
2. calculate working responses $z_i^{(t)}$ and weights $w_i^{(t)}$
3. calculate $\boldsymbol{\beta}^{(t+1)}$ by weighted least squares
4. repeat steps 2 and 3 till convergence.

Introduction to GLM's: Standard errors

For the MLE $\hat{\boldsymbol{\beta}}$ it holds that (asymptotically)

$$\hat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, \phi(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1})$$

Thus, standard errors of the estimators $\hat{\beta}_j$ are the respective diagonal elements of the estimated variance/covariance matrix

$$\widehat{\text{var}(\hat{\boldsymbol{\beta}})} = \phi(\mathbf{X}^\top \hat{\mathbf{W}} \mathbf{X})^{-1}$$

with $\hat{\mathbf{W}} = \mathbf{W}(\hat{\boldsymbol{\mu}})$. Note that $(\mathbf{X}^\top \hat{\mathbf{W}} \mathbf{X})^{-1}$ is a by-product of the last IWLS iteration. If ϕ is unknown, an estimator is required.

Introduction to GLM's: Dispersion estimator

There are practical difficulties when estimating ϕ by ML.
A **method-of-moments** like estimator is developed considering the ratios

$$\phi = \frac{\mathrm{E}(y_i - \mu_i)^2}{V(\mu_i)}\,, \qquad \text{for all } i = 1, \ldots, n$$

Averaging over all these ratios and assuming that the μ_i's are known results in the estimator

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{V(\mu_i)}$$

However, since $\boldsymbol{\beta}$ is unknown we better use the bias-corrected version (also known as the mean generalized Pearson's chi-square statistic)

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \frac{1}{n-p}X^2$$

The glm Function

Generalized linear models can be fitted in R using the glm function, which is similar to lm for fitting linear models. The arguments to a glm call are as follows:

```
glm(formula, family = gaussian, data, weights, subset,
   na.action, start = NULL, etastart, mustart, offset,
   control = glm.control(...), model = TRUE,
   method = "glm.fit", x = FALSE, y = TRUE,
   contrasts = NULL, ...)
```

The glm Function

Formula argument:

The formula is specified for a glm as e.g.

```
    y ~ x1 + x2
```

where x1 and x2 are the names of

- numeric vectors (continuous predictors)
- factors (categorial predictors)

All the variables used in the formula must be in the workspace or in the data frame passed to the data argument.

The `glm` Function

Formula argument:

Other symbols that can be used in the formula are:

- `a:b` for the interaction between `a` and `b`
- `a*b` which expands to `1 + a + b + a:b`
- `.` first order terms of all variables in `data`
- `-` to exclude a term (or terms)
- `1` intercept (default)
- `-1` without intercept

The `glm` Function

Family argument:

The family argument defines the response distribution (**variance** function) and the **link** function. The exponential family functions available in R are e.g.

- `gaussian(link = "identity")`
- `binomial(link = "logit")`
- `poisson(link = "log")`
- `Gamma(link = "inverse")`

The glm Function

Extractor functions:

The glm function returns an object of class c("glm", "lm"). There are several methods available to access or display components of a glm object, e.g.

- residuals()
- fitted()
- predict()
- coef()
- deviance()
- summary()
- plot()

The glm Function: Example

Refit **life expectancies** model using glm().
The first part contains the same information as from lm()

```
> mod<-glm(life.expectancy ~ urban+log(physicians)+temperature)
> summary(mod)

Call:
glm(formula=life.expectancy ~ urban+log(physicians)+temperature)

Deviance Residuals:
    Min        1Q    Median       3Q      Max
-14.033    -3.089     0.379    3.328   12.144

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     66.70367    1.79065  37.251  < 2e-16 ***
urban            8.76445    2.53243   3.461 0.000711 ***
log(physicians)  3.51370    0.39341   8.931 1.97e-15 ***
temperature     -0.03008    0.05668  -0.531 0.596408
```

The `glm` Function: Example

Since the default `family="gaussian"`, deviance residuals corresponds to ordinary residuals as in a linear model.

A five-number summary of those raw residuals is given.

Wald tests

Remember that for the MLE it asymptotically holds that

$$\hat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, \phi(\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1})$$

Thus, we can utilize this to construct a test statistic on the significance of a coefficient, say β_j for $j = 1, \ldots, p - 1$.
If we test

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0$$

we can use the test statistic

$$t = \frac{\hat{\beta}_j}{\sqrt{\hat{\phi}(\mathbf{X}^\top \hat{\mathbf{W}}\mathbf{X})^{-1}_{j+1,j+1}}}$$

which under H_0 asymptotically follows a t distribution with $n - p$ degrees of freedom.

The `glm` Function: Example

The second part contains some new information on estimated **dispersion** and **goodness-of-fit aspects** which we will discuss later in detail.

First the dispersion estimate (if necessary) $\hat{\phi}$ is provided

```
(Dispersion parameter for gaussian family taken to be 22.9815)
```

This estimate is simply the squared residual standard error (that was 4.794 in the `summary(lm())`).

(Scaled) Deviance

Next there is the **deviance** of two models and the number of missing observations:

```
    Null deviance: 11109.6  on 145  degrees of freedom
Residual deviance:  3263.4  on 142  degrees of freedom
  (23 observations deleted due to missingness)
```

The first refers to the **null model** which corresponds to a model with intercept only (the iid assumption, no explanatory variables). The associated degrees of freedom are $n-1$.

The second refers to our **fitted model** with $p-1$ explanatory variables in the predictor and, thus, with associated degrees of freedom $n-p$.

(Scaled) Deviance

The **deviance** of a model is defined as the distance of log-likelihoods, i.e.

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2\phi \left(\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) - \ell(\mathbf{y}|\mathbf{y})\right)$$

Here, $\hat{\boldsymbol{\mu}}$ are the fitted values under the considered model (maximizing the log-likelihood under the given parametrization), and $\mathbf{y}$ denote the estimated means under a model without any restriction at all (thus $\hat{\boldsymbol{\mu}} = \mathbf{y}$ in such a **saturated model**).

(Scaled) Deviance

For any member of the LEF the deviance equals

$$
\begin{aligned}
D(\mathbf{y}, \hat{\boldsymbol{\mu}}) &= -2\phi \sum_{i=1}^{n} \frac{(y_i\hat{\theta}_i - y_i\tilde{\theta}_i) - (b(\hat{\theta}_i) - b(\tilde{\theta}_i))}{\phi} \\
&= -2 \sum_{i=1}^{n} \left\{(y_i\hat{\theta}_i - y_i\tilde{\theta}_i) - (b(\hat{\theta}_i) - b(\tilde{\theta}_i))\right\}
\end{aligned}
$$

where $\tilde{\theta}_i$ denotes the estimate of θ_i under the saturated model. Under the saturated model, there are as many mean parameters μ_i allowed as observations y_i.

Note that for LEF members the **deviance**

$$
D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2 \sum_{i=1}^{n} \left\{(y_i\hat{\theta}_i - y_i\tilde{\theta}_i) - (b(\hat{\theta}_i) - b(\tilde{\theta}_i))\right\}
$$

doesn't depend on the dispersion!

(Scaled) Deviance

Example: Gaussian responses ($\phi = \sigma^2$) with identity link (LM)

$$\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\sigma^2}$$
$$\ell(\mathbf{y}|\mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2)$$

Therefore the deviance equals the **sum of squared errors**, i.e.

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2\phi\left(\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) - \ell(\mathbf{y}|\mathbf{y})\right) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \mathrm{SSE}(\hat{\boldsymbol{\beta}})$$

(Scaled) Deviance

Finally we have

```
AIC: 877.94

Number of Fisher Scoring iterations: 2
```

The **Akaike Information Criterion (AIC)** also assess the fit penalizing for the total number of parameters $p+1$ (linear predictor and dispersion in this case) and is defined as

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) + 2(p+1)$$

The smaller the AIC value the better the fit. Use AIC only to compare different models (not necessarily nested).
Sometimes, the term $-2\ell(\hat{\boldsymbol{\mu}}|\mathbf{y})$ is called **disparity**.

Residuals

Several different ways to define residuals in a GLM:

```
residuals(object, type = c("deviance", "pearson", "working",
                           "response", "partial"), ...)
```

- `deviance`: write deviance as $\sum_{i=1}^{n} d(y_i, \hat{\mu}_i)^2$
- `pearson`: $r_i^P = (y_i - \hat{\mu}_i)/\sqrt{V(\hat{\mu}_i)}$
- `working`: $r_i^W = \hat{z}_i - \hat{\eta}_i = (y_i - \hat{\mu}_i)g'(\hat{\mu}_i)$ (remember that $g'(\hat{\mu}_i) = 1/V(\hat{\mu}_i)$ for canonical link models)
- `response`: $y_i - \hat{\mu}_i$
- `partial`: $r_i^P + \hat{\beta}_j x_{ij}$ is the partial residual for the j-th covariate

Except the partial residuals, these types are all equivalent for LM's.

Residuals

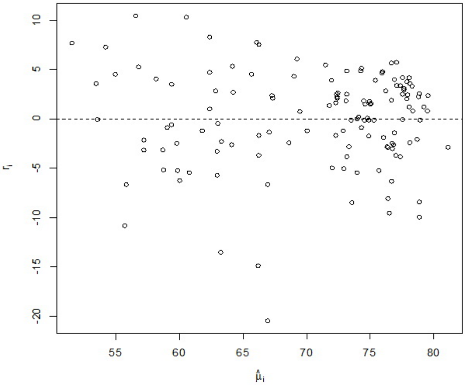
Deviance residuals are the default used in R since they reflect the same criterion as used in the fitting.

Plot deviance residuals against fitted values:

```
> plot(residuals(mod) ~ fitted(mod),
+ xlab = expression(hat(mu)[i]),
+ ylab = expression(r[i]))
> abline(0, 0, lty = 2)
```

Residuals

Deviance/Pearson/response/working residuals vs. fitted values:



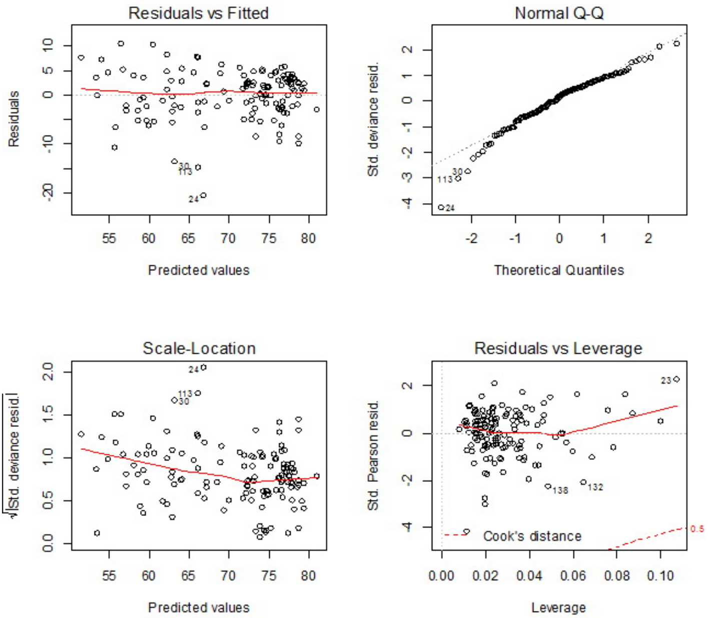
The `glm` Function: Plot

The `plot()` function gives the following sequence of plots:

- deviance residuals vs. fitted values
- Normal Q-Q plot of deviance residuals standardized to unit variance
- scale-location plot of standardized deviance residuals
- standardized deviance residuals vs. leverage with Cook's distance contours

```
> plot(mod)
```

The `glm` Function: Plot



Black Cherry Trees Revisited

So far we considered (Box-Cox transformation) models like

- $V_i^{1/3} \overset{ind}{\sim} \text{Normal}(\mu_i, \sigma^2)$, $\text{E}(V^{1/3}) = \mu = H + D$
- $\log(V_i) \overset{ind}{\sim} \text{Normal}(\mu_i, \sigma^2)$, $\text{E}(\log(V)) = \mu = \log(H) + \log(D)$

In what follows we will assume that a GLM holds with

- $V_i \overset{ind}{\sim} \text{Normal}(\mu_i, \sigma^2)$ and $g(\text{E}(V)) = \eta$.

More specifically, we like to check out the models:

- $\mu^{1/3} = H + D$
- $\log(\mu) = \log(H) + \log(D)$.

These models on the **observations scale** can be easily fitted using `glm()`.

Black Cherry Trees Revisited

$$V_i \overset{ind}{\sim} \text{Normal}(\mu_i, \sigma^2),\ \mu^{1/3} = H + D$$

```
> pmodel <- glm(V ~ H + D, family = gaussian(link=power(1/3)))
> summary(pmodel)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.051322   0.224095  -0.229 0.820518
H            0.014287   0.003342   4.274 0.000201 ***
D            0.150331   0.005838  25.749  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 6.577063)

    Null deviance: 8106.08  on 30  degrees of freedom
Residual deviance:  184.16  on 28  degrees of freedom
AIC: 151.21

Number of Fisher Scoring iterations: 4
```

Black Cherry Trees Revisited

$$V_i \overset{ind}{\sim} \text{Normal}(\mu_i, \sigma^2),\ \mu^{1/3} = H + D$$

```
> AIC(pmodel)
[1] 151.2102
> -2*logLik(pmodel) + 2*4
'log Lik.' 151.2102 (df=4)

> logLik(pmodel)
'log Lik.' -71.60508 (df=4)
> sum(log(dnorm(V,pmodel$fit,sqrt(summary(pmodel)$disp*28/31))))
[1] -71.60508

> sum(residuals(pmodel)^2)
[1] 184.1577
> deviance(pmodel)
[1] 184.1577
> sum((V-mean(V))^2) # Null Deviance
[1] 8106.084
```

Black Cherry Trees Revisited

$$V_i \overset{ind}{\sim} \text{Normal}(\mu_i, \sigma^2), \; \log(\mu) = \log(H) + \log(D)$$

```
> summary(glm(V ~ log(H) + log(D), family = gaussian(link=log)))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.53700    0.94352  -6.928 1.57e-07 ***
log(H)       1.08765    0.24216   4.491 0.000111 ***
log(D)       1.99692    0.08208  24.330  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 6.41642)

    Null deviance: 8106.08  on 30  degrees of freedom
Residual deviance:  179.66  on 28  degrees of freedom
AIC: 150.44

Number of Fisher Scoring iterations: 4
```

Gamma Regression

Gamma responses: $y \sim \text{Gamma}(a, \lambda)$ with density function

$$f(y|a,\lambda) = \exp(-\lambda y)\lambda^a y^{a-1}\frac{1}{\Gamma(a)}, \qquad a, \lambda, y > 0$$

with $\text{E}(y) = a/\lambda$ and $\text{var}(y) = a/\lambda^2$.

Mean parametrization needed!

Gamma Regression

Reparametrization: define $\mu = \nu/\lambda, \;\; \nu = a$

$$f(y|a,\lambda) = \exp(-\lambda y)\lambda^{a} y^{a-1}\frac{1}{\Gamma(a)}$$

$$\begin{aligned} f(y|\mu,\nu) &= \exp\left(-\frac{\nu}{\mu}y\right)\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\frac{1}{\Gamma(\nu)} \\ &= \exp\left(\frac{y\left(-\frac{1}{\mu}\right) - \log\mu}{1/\nu} + \nu\log\nu + (\nu-1)\log y - \log\Gamma(\nu)\right) \end{aligned}$$

LEF member with:

$\theta = -1/\mu, \;\; b(\theta) = \log\mu = -\log(-\theta), \;\;$ and $\phi = 1/\nu$.

Gamma Regression

Gamma(μ, ν) belongs to the **LEF** with

$$\theta = -1/\mu, \quad b(\theta) = \log \mu = -\log(-\theta), \quad \phi = 1/\nu.$$

Thus,

$$\begin{aligned} \mathrm{E}(y) &= b'(\theta) = -\frac{-1}{-\theta} = -\frac{1}{\theta} = \mu \\ \mathrm{var}(y) &= \phi b''(\theta) = \phi \frac{1}{\theta^2} = \phi\mu^2 \end{aligned}$$

with dispersion $\phi = 1/\nu$ and variance function $V(\mu) = \mu^2$.

Coefficient of variation:

$$\frac{\sqrt{\mathrm{var}(y_i)}}{\mathrm{E}(y_i)} = \frac{\sqrt{\phi\mu_i^2}}{\mu_i} = \sqrt{\phi} = \text{constant for all } i = 1, \ldots, n.$$

Gamma Regression

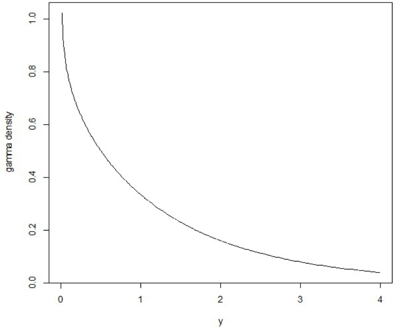
Form of the Gamma(μ, ν) density function is determined by ν.
Functions in R are based on `shape` $(= 1/\phi)$ and `scale` $(= \phi\mu)$

```
> y <- (1:400)/100
> shape <- 0.9
> scale <- 1.5
> plot(y, dgamma(y, shape=shape, scale=scale))

> mean(rgamma(10000, shape=shape, scale=scale)); shape*scale
[1] 1.374609
[1] 1.35
> var(rgamma(10000, shape=shape, scale=scale)); shape*(scale)^2
[1] 2.001009
[1] 2.025
```

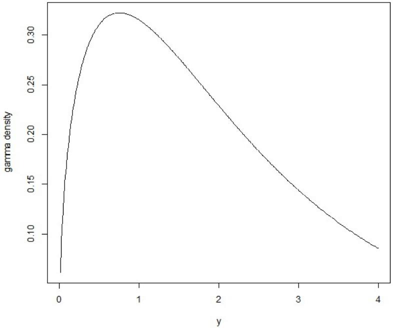
Gamma Regression

Gamma distributions are generally **skewed to the right**.

`shape` < 1 (0.9 left)



`shape` > 1 (1.5 right)



Special cases: $\nu = 1/\phi = 1$ (exponential) and $\nu \to \infty$ (normal)

Gamma Regression: Link Function

What's an **appropriate link** function?

- Canonical link function: $\eta = \theta = -\frac{1}{\mu}$ (**inverse-link**). Since we need $\mu > 0$ we need $\eta < 0$ giving complicated restriction on $\boldsymbol{\beta}$.
- Thus, the **log-link** is often used without restrictions on η, i.e.

$$\log \mu = \eta$$

Gamma Regression: Deviance

Assume that $y_i \sim \text{Gamma}(\mu_i, \phi)$ (independent) and $\log \mu_i = \eta_i$. Then

$$\ell(\hat{\boldsymbol{\mu}}, \phi | \mathbf{y}) = \sum_{i=1}^{n} \left\{ \frac{y_i \left(-\frac{1}{\hat{\mu}_i}\right) - \log \hat{\mu}_i}{\phi} + c(y_i, \phi) \right\}$$

$$\ell(\mathbf{y}, \phi | \mathbf{y}) = \sum_{i=1}^{n} \left\{ \frac{y_i \left(-\frac{1}{y_i}\right) - \log y_i}{\phi} + c(y_i, \phi) \right\}$$

and thus the **scaled deviance** equals

$$\begin{aligned} \frac{1}{\phi} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) &= -\frac{2}{\phi} \sum_{i=1}^{n} \left\{ \left(-\frac{y_i}{\hat{\mu}_i} - \log \hat{\mu}_i \right) - (-1 - \log y_i) \right\} \\ &= -\frac{2}{\phi} \sum_{i=1}^{n} \left\{ \log \frac{y_i}{\hat{\mu}_i} - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right\} \end{aligned}$$

Gamma Regression: Dispersion

Method of moments is used to estimate the dispersion parameter. We have a sample $y_1, \ldots, y_n$ with

$$E(y_i) = \mu_i \quad \text{and} \quad \text{var}(y_i) = \phi \mu_i^2, \qquad i = 1, \ldots, n$$

Consider $z_i = y_i/\mu_i$ with $E(z_i) = 1$ and $\text{var}(z_i) = \phi$ (z_i are iid). Thus,

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \left(\frac{y_i}{\hat{\mu}_i} - 1 \right)^2 = \frac{1}{n-p} \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right)^2$$

which is equivalent to the mean **Pearson** statistic.

The `glm` Function: Example Life Expectancy

We now assume that life expectancy follows a **gamma** model.

```
> gmod<-glm(life.expectancy~urban+log(physicians)+temperature,
+             family=Gamma(link="log"))
> summary(gmod)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.2020227  0.0269393 155.981  < 2e-16 ***
urban            0.1110928  0.0380990   2.916  0.00412 **
log(physicians)  0.0543425  0.0059186   9.182 4.61e-16 ***
temperature     -0.0002702  0.0008527  -0.317  0.75180
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Given `urban` and `log(physicians)` are already in the model, `temperature` seems to be again **irrelevant** as an additional predictor.

The `glm` Function: Example Life Expectancy

The next part of the output contains information about:

```
(Dispersion parameter for Gamma family taken to be 0.005201521)
```

The dispersion estimate $\hat{\phi}$ is the mean Pearson statistic

```
> # direct from summary(.)
> summary(gmod)$dispersion
[1] 0.005201521
> # or explicitly calculated as
> sum(residuals(gmod, type="pearson")^2)/gmod$df.resid
[1] 0.005201521
```

giving the estimated response variance as $\widehat{\text{var}}(y_i) = 0.0052 V(\hat{\mu}_i)$.

The `glm` Function: Example Life Expectancy

```
(Dispersion parameter for Gamma family taken to be 0.005201521)

    Null deviance: 2.42969  on 145  degrees of freedom
Residual deviance: 0.76096  on 142  degrees of freedom
  (23 observations deleted due to missingness)
AIC: 896.14

Number of Fisher Scoring iterations: 4
```

For the scaled deviance we get

$$\frac{1}{\hat{\phi}} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \frac{0.76096}{0.00520} = 146.2957$$

which is pretty close its associated degrees of freedom 142.

The `glm` Function: Example Life Expectancy

Residual Deviance Test:

Model $(*)$: $y_i \overset{ind}{\sim} \text{Gamma}(\mu_i = \exp(\eta_i), \phi)$, $i = 1, \ldots, n$.

Reject model $(*)$ at level α if

$$\frac{1}{\phi} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) > \chi^2_{1-\alpha, n-p}$$

Since the dispersion ϕ is unknown, we use its estimate $\hat{\phi}$ instead and reject model $(*)$ if

$$\frac{1}{\hat{\phi}} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) > \chi^2_{1-\alpha, n-p}$$

```
> 1-pchisq(deviance(gmod)/summary(gmod)$disp, gmod$df.resid)
[1] 0.3852  # p-value
```

The `glm` Function: Example Life Expectancy

Partial Deviance Test:

Consider the model $g(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$ with $\dim(\boldsymbol{\beta}_1) = p_1$, $\dim(\boldsymbol{\beta}_2) = p_2$ and $p = p_1 + p_2$. Now calculate

- $\hat{\boldsymbol{\mu}}_1 = g^{-1}(\mathbf{X}_1\hat{\boldsymbol{\beta}}_1)$: the fitted means under the reduced model with design $\mathbf{X}_1$ only (corresponds to $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$)
- $\hat{\boldsymbol{\mu}}_2 = g^{-1}(\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2)$: the fitted means under the full model with design $\mathbf{X}_1$ and $\mathbf{X}_2$
- $\hat{\phi} = X^2/(n-p)$: dispersion estimate under the full model

Reject H_0 at level α if

$$\frac{(D(\mathbf{y}, \hat{\boldsymbol{\mu}}_1) - D(\mathbf{y}, \hat{\boldsymbol{\mu}}_2))/p_2}{\hat{\phi}} > F_{1-\alpha, p_2, n-p}$$

The `glm` Function: Example Life Expectancy

Reject $H_0 : \beta_{\text{temp}} = 0$ if

$$\frac{(D(\mathbf{y}, \hat{\boldsymbol{\mu}}_1) - D(\mathbf{y}, \hat{\boldsymbol{\mu}}_2))/1}{\hat{\phi}} > F_{1-\alpha,1,n-p}$$

```
> (dev2 <- deviance(gmod))
[1] 0.7609569
> (hatphi <- sum(residuals(gmod, type="pearson")^2)/gmod$df.r)
[1] 0.005201521

> gmod1 <- glm(life.exp ~ urban + log(physicians),
+                family=Gamma(link="log"))
> (dev1 <- deviance(gmod1))
[1] 0.761484

> (F <- ((dev1-dev2)/1)/hatphi)
[1] 0.1013431
> 1-pf(F, 1, gmod$df.r)
[1] 0.7506915
```

The glm Function: Example Life Expectancy

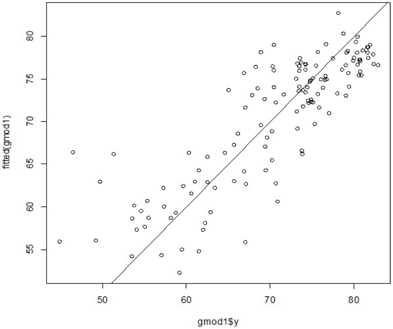
ANalysis Of deViAnce (ANOVA):
Much easier to use again anova():

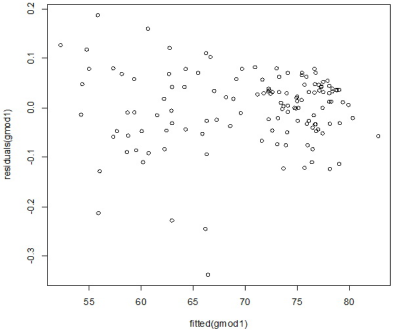
```
> anova(gmod, test="F")
Analysis of Deviance Table

Model: Gamma, link: log
Response: life.expectancy
Terms added sequentially (first to last)

                Df Deviance Resid. Df Resid. Dev      F Pr(>F)
NULL                              145    2.42969
urban            1  1.09627       144    1.33342 210.76 <2e-16 **
log(physicians) 1  0.57194       143    0.76148 109.96 <2e-16 **
temperature      1  0.00053       142    0.76096   0.10 0.7507
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

The glm Function: Example Life Expectancy

Some Diagnostic Plots:

```
> plot(gmod1$y, fitted(gmod1), xlim=c(45,85), ylim=c(45,85))
> abline(0,1)

> plot(fitted(gmod1), residuals(gmod1))
```





The `glm` Function: Example Life Expectancy

Something about the usage of `predict()`

```
predict(object, newdata = NULL,
        type = c("link", "response", "terms"),
        se.fit = FALSE, dispersion = NULL, ...)
```

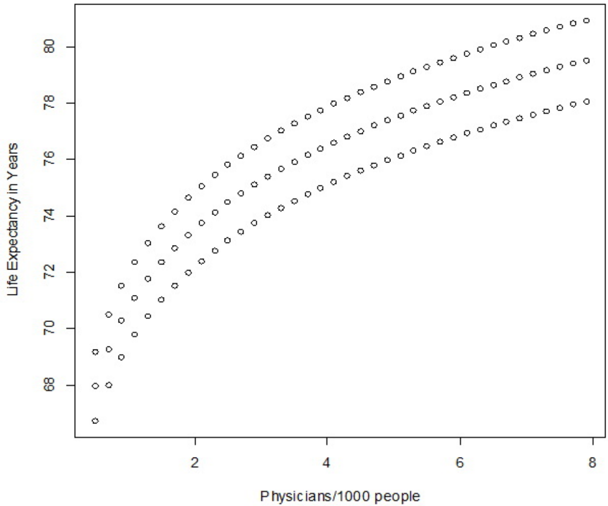
- `newdata`: data frame with predictor values for which to predict.
- `type`: default is on the scale of the linear predictors. The ”terms” option returns a matrix giving the fitted values of each term in the model formula on the linear predictor scale.
- `se.fit`: logical indicator if standard errors are required.
- `dispersion`: parameter value used in computing standard errors (if omitted, that returned by summary).

The glm Function: Example Life Expectancy

Predict life expectancy for urbanization rates of 37, 56, and 74 % (the empirical 25, 50, and 75 % data quartiles).

```
> u.q <- quantile(urban, probs = seq(0.25, 0.75, 0.25),
+                 na.rm="TRUE")
> new <- expand.grid(physicians=seq(0.5, 8, 0.2), urban = u.q)

> p <- predict(gmod1, newdata=new, type="response")

> plot(new$physicians, p, xlab="Physicians/1000 people",
+      ylab="Life Expectancy in Years")
```

The glm Function: Example Life Expectancy



The `glm` Function: Example Life Expectancy

Remarks about other predictions:

```
> # predict linear predictor \hat\eta_i
> pl <- predict(gmod1, newdata=new, type="link")
       1        2        3 ...
4.202555 4.221124 4.234993 ...

> # predict each term in the linear predictor separately
> pt <- predict(gmod1, newdata=new, type="terms")
        urban log(physicians)
1 -0.01994721    -0.023863823
2 -0.01994721    -0.005295193
3 -0.01994721     0.008573900
:
attr(,"constant")
[1] 4.246366
> attr(pt, "const") + pt[ ,"urban"] + pt[ ,"log(physicians)"]
       1        2        3 ...
4.202555 4.221124 4.234993 ...
```

Logistic Regression

Response Variables y_i, $i = 1, \ldots, n$:

- **ungrouped**: each variable y_i can take one of two values, say success/failure (or 0/1),
- **grouped**: the variable $m_i y_i$ is the number of successes in a given number of m_i trials; y_i is the **relative** success frequency, $m_i y_i$ denotes the **absolute** success frequency.

Both situations correspond to a **Binomial**(m_i, π_i) model, where in the ungrouped case we have $m_i = 1$.

Question: Is the binomial distribution also a member of the **linear exponential family (LEF)?**

Logistic Regression: LEF Member

Standardized Binomial: $my \sim \text{Binomial}(m, \pi)$ (m known)

$$
\begin{aligned}
f(y|m,\pi) = \Pr(Y = y) = \Pr(mY = my) &= \binom{m}{my} \pi^{my} (1-\pi)^{m-my} \\
&= \exp\left(\log \binom{m}{my} + my \log \pi + m(1-y)\log(1-\pi) \right) \\
&= \exp\left(\frac{y \log \frac{\pi}{1-\pi} - \log \frac{1}{1-\pi}}{1/m} + \log \binom{m}{my} \right), \quad y = 0, \frac{1}{m}, \frac{2}{m}, \ldots, 1.
\end{aligned}
$$

If m is another unknown parameter, this is no longer a LEF member!

Logistic Regression: LEF Member

Standardized Binomial: $my \sim \text{Binomial}(m, \pi)$ (m known)

$$f(y|m,\pi) = \exp\left(\frac{y \log \frac{\pi}{1-\pi} - \log \frac{1}{1-\pi}}{1/m} + \log \binom{m}{my} \right), \quad y = 0, \frac{1}{m}, \frac{2}{m}, \dots, 1.$$

Let $\theta = \log \frac{\pi}{1-\pi}$, $(\pi = e^{\theta}/(1+e^{\theta}))$ and $\phi = 1$ then we have identified another LEF member with

$$a = \frac{1}{m}, \quad b(\theta) = \log \frac{1}{1-\pi} = \log(1+\exp(\theta)), \quad c(y,\phi) = \log \binom{m}{my}.$$

Notice: the **dispersion** parameter $\phi = 1$ is **known** in this case and $a = 1/m$ is a **weight** and considered to be **fixed**!

Logistic Regression: Link

For a sample $m_i y_i \overset{ind}{\sim} \text{Binomial}(m_i, \pi_i)$, $y_i = 0, 1/m_i, \ldots, 1$, we have $\text{E}(m_i y_i) = m_i \pi_i$ and $\text{var}(m_i y_i) = m_i \pi_i (1 - \pi_i)$ and thus

$$\text{E}(y_i) = \pi_i =: \mu_i \qquad \text{and} \qquad \text{var}(y_i) = \frac{1}{m_i} \mu_i (1 - \mu_i)$$

with restriction $0 < \mu_i < 1$.

Canonical link $g(\mu_i) = b'^{-1}(\mu_i) = \theta_i$ is the **logit link**

$$\text{logit}(\mu_i) = \log \frac{\mu_i}{1 - \mu_i} = \log \frac{m_i \mu_i}{m_i - m_i \mu_i} = \theta_i = \eta_i$$

$$\Rightarrow \quad \mu_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \,.$$

However, in principal any inverse of a continuous distribution function can be used as $g(\cdot)$.

Logistic Regression: Link

The name **logit** refers to the distribution function of a logistic distributed random variable with density function

$$f(y|\mu,\tau) = \frac{\exp((y-\mu)/\tau)}{\tau\Big(1+\exp((y-\mu)/\tau)\Big)^2}\,, \qquad \mu \in \mathbb{R},\ \tau > 0\,,$$

for which $\mathrm{E}(y) = \mu$ and $\mathrm{var}(y) = \tau^2\pi^2/3$ holds.

The density and the cdf of its standard form ($\mu = 0$, $\tau = 1$) is

$$f(y|0,1) = \frac{\exp(y)}{\Big(1+\exp(y)\Big)^2}\,, \quad y \in \mathbb{R}, \qquad F(y|0,1) = \frac{\exp(y)}{1+\exp(y)}$$

for which $\mathrm{E}(y) = 0$ and $\mathrm{var}(y) = \pi^2/3$ holds.

$F(y|0,1)$ corresponds to the inverse logit link.

Logistic Regression: Links

With $g^{-1}(\eta) = \Phi(\eta)$ we refer to a **probit model**. Logit- and probit link are both **symmetric** links.

Extreme value distribution:
Maximum

$$F_{max}(y) = \exp(-\exp(-y))\,, \qquad y \in \mathbb{R}$$

with $\mathrm{E}(y) = \gamma$ (Euler constant $\gamma = 0.577216$) and $\mathrm{var}(y) = \pi^2/6$. The inverse of $F_{max}(\cdot)$ results in the **log-log link** and equals

$$g(\mu) = -\log(-\log(\mu)).$$

Logistic Regression: Links

Minimum

$$F_{min}(y) = 1 - F_{max}(-y) = 1 - \exp(-\exp(y))\,, \qquad y \in \mathbb{R}$$

with $\mathrm{E}(y) = -\gamma$ and $\mathrm{var}(y) = \pi^2/6$.
The inverse of $F_{min}(\cdot)$ is called **complementary log-log link** and equals $g(\mu) = \log(-\log(1-\mu))$.

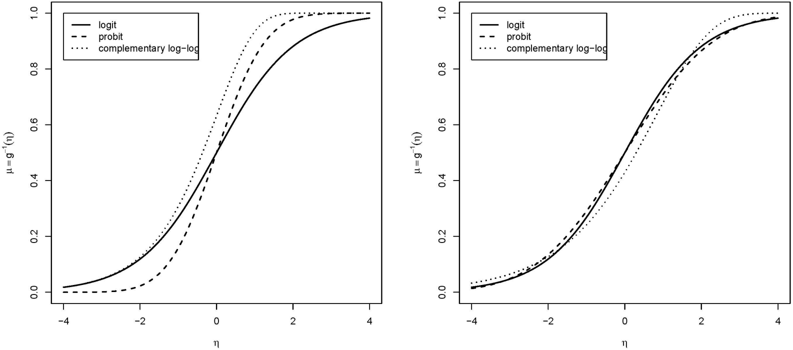
Both extreme value distribution functions give **asymmetric** links.

Logistic Regression: Links

R allows for `family=binomial` to use several specifications of the link function: `logit`, `probit`, `cauchit`, as also `log` and `cloglog`.

```
> euler <- 0.577216
> mu.logit  <-function(eta) 1/(1 + exp(-eta))
> mu.probit <-function(eta) pnorm(eta, 0, pi/sqrt(3))
> mu.cloglog<-function(eta) 1-exp(-exp(-euler+eta/sqrt(2)))
> plot(mu.logit, (-4): 4, xlim = c(-4, 4), ylim = c(0,1),
+      xlab = expression(eta),
´+     ylab = expression(mu == g^-1 * (eta)), lwd=2)
> curve(mu.probit,  (-4):4, add = TRUE, lty = 2, lwd=2)
> curve(mu.cloglog, (-4):4, add = TRUE, lty = 3, lwd=2)
> legend(-4, 1, c("logit", "probit", "complementary log-log"),
+        lty = 1:3, lwd=2)
```

Logistic Regression: Links



Logistic Regression: Deviance

For $m_i y_i \sim \text{Binomial}(m_i, \mu_i)$ (m_i known) we write the ith log-likelihood contribution as

$$\log f(y_i | m_i, \mu_i) = m_i y_i \log \frac{\mu_i}{1-\mu_i} - m_i \log \frac{1}{1-\mu_i} + \log \binom{m_i}{m_i y_i}$$

to get the sample (model and saturated) log-likelihood functions

$$\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) = \sum_{i=1}^{n} \left\{ m_i y_i \log \frac{\hat{\mu}_i}{1-\hat{\mu}_i} - m_i \log \frac{1}{1-\hat{\mu}_i} + \log \binom{m_i}{m_i y_i} \right\}$$

$$\ell(\mathbf{y}|\mathbf{y}) = \sum_{i=1}^{n} \left\{ m_i y_i \log \frac{y_i}{1-y_i} - m_i \log \frac{1}{1-y_i} + \log \binom{m_i}{m_i y_i} \right\} .$$

Logistic Regression: Deviance

$$\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) = \sum_{i=1}^{n} \left\{ m_i y_i \log \frac{\hat{\mu}_i}{1-\hat{\mu}_i} - m_i \log \frac{1}{1-\hat{\mu}_i} + \log \binom{m_i}{m_i y_i} \right\}$$
$$\ell(\mathbf{y}|\mathbf{y}) = \sum_{i=1}^{n} \left\{ m_i y_i \log \frac{y_i}{1-y_i} - m_i \log \frac{1}{1-y_i} + \log \binom{m_i}{m_i y_i} \right\} .$$

Because of $\phi = 1$ and $a_i = 1/m_i$ the resulting (scaled) deviance is

$$\frac{1}{\phi} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2 \sum_{i=1}^{n} \left\{ m_i y_i \left(\log \frac{\hat{\mu}_i}{y_i} + \log \frac{1-y_i}{1-\hat{\mu}_i} \right) - m_i \log \frac{1-y_i}{1-\hat{\mu}_i} \right\}$$
$$= 2 \sum_{i=1}^{n} m_i \left\{ (1-y_i) \log \frac{1-y_i}{1-\hat{\mu}_i} + y_i \log \frac{y_i}{\hat{\mu}_i} \right\} .$$

Notice: for $y_i = 0$ or 1 independent of $\hat{\mu}_i$ (because $x \log x = 0$ for $x = 0$) the respective term in the deviance component disappears.

Logistic Regression: Deviance

For binary data $y_i \in \{0, 1\}$ ($m_i = 1$ for all i) we get

$$\ell(\mu_i|y_i) = \begin{cases} \log(1-\mu_i) & \text{if} \quad y_i = 0\,, \\ \log \mu_i & \text{if} \quad y_i = 1 \end{cases}$$

and

$$d(y_i, \hat{\mu}_i) = \begin{cases} -2\log(1-\hat{\mu}_i) & \text{if} \quad y_i = 0\,, \\ -2\log \hat{\mu}_i & \text{if} \quad y_i = 1\,. \end{cases}$$

The deviance increment $d(y_i, \hat{\mu}_i)$ describes the fraction of a binary response of the maximized sample log-likelihood function

$$\ell(\hat{\boldsymbol{\mu}}|\mathbf{y}) = \sum_{i=1}^{n} \ell(\hat{\mu}_i|y_i) = -\frac{1}{2}\sum_{i=1}^{n} d(y_i, \hat{\mu}_i)\,.$$

Logistic Regression: Tolerance Distribution

Bioassay: experimental study based on binary responses, e.g. testing the effect of various concentrations in animal experiments.

Number of animals responding is considered as binomial response.

Example: Insecticide applied on groups (**batches**) of insects of known sizes. When applying a low dose to a group, then no insect will probably fall out. If a high dose is given to another group, many insects of this group will die.

If an insect dies or not when receiving a certain dosage depends on the **tolerance** of the animal. Insects with a low tolerance will rather die on a certain dose than any other with a high tolerance.

Logistic Regression: Tolerance Distribution

Assumption: the tolerance U of an insect is a random variable with density $f(u)$. Insects with tolerance $U < d_i$ will die.

Probability that an animal dies when receiving dose d_i is

$$p_i = \Pr(U < d_i) = \int_{-\infty}^{d_i} f(u)\, du\,.$$

If $U \sim \textbf{Normal}(\mu, \sigma^2)$, then

$$p_i = \Phi\left(\frac{d_i - \mu}{\sigma}\right)\,.$$

With $\beta_0 = -\mu/\sigma$ and $\beta_1 = 1/\sigma$ this gives

$$p_i = \Phi\left(\beta_0 + \beta_1 d_i\right) \quad \text{or} \quad \text{probit}(p_i) = \Phi^{-1}(p_i) = \beta_0 + \beta_1 d_i\,,$$

i.e. a **probit model** for mortality p_i depending on the dose d_i.

Logistic Regression: Tolerance Distribution

If U follows a **logistic**(μ, τ) model then

$$p_i = \Pr(U \leq d_i) = \int_{-\infty}^{d_i} \frac{\exp((u-\mu)/\tau)}{\tau\Big(1+\exp((u-\mu)/\tau)\Big)^2}\, du$$

$$= \frac{\exp((d_i-\mu)/\tau)}{1+\exp((d_i-\mu)/\tau)}\,.$$

With $\beta_0 = -\mu/\tau$ and $\beta_1 = 1/\tau$ we get

$$p_i = \frac{\exp(\beta_0 + \beta_1 d_i)}{1+\exp(\beta_0+\beta_1 d_i)} \quad \text{or} \quad \text{logit}(p_i) = \beta_0 + \beta_1 d_i$$

giving a **logistic link model** for p_i.

Logistic Regression: Tolerance Distribution

Example: Effect of poison given to the *Tobacco Budworm*. Groups of 20 moths of both `sex` are exposed to various doses of a poison and the number of killed animals has been recorded.

	Dose in μg					
sex	1	2	4	8	16	32
male	1	4	9	13	18	20
female	0	2	6	10	12	16

Logistic Regression: Tolerance Distribution

Doses are powers of 2. Thus, we use `ldose = ` $\log_2(\text{dose})$ as predictor variable.

```
> (ldose <- rep(0:5, 2))
 [1] 0 1 2 3 4 5 0 1 2 3 4 5

> (sex <- factor(rep(c("M", "F"), c(6, 6))))
 [1] M M M M M M F F F F F F
Levels: F M

> (dead <- c(1,4,9,13,18,20,0,2,6,10,12,16))
 [1]  1  4  9 13 18 20  0  2  6 10 12 16
```

Logistic Regression: Tolerance Distribution

- Specification of binomial responses in R by means of a matrix SF (success/failure), in which the **first** (second) column contains the number of **successes** (failures).

- Model describes the **probability of success** (the number of killed animals in our case) at a certain dosage.

```
> (SF <- cbind(dead, alive = 20-dead))
      dead alive
 [1,]    1    19
 [2,]    4    16
   :
[12,]   16     4
```

Logistic Regression: Tolerance Distribution

```
> summary(budworm.lg <- glm(SF ~ sex*ldose, family = binomial))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9935     0.5527  -5.416 6.09e-08 ***
sexM          0.1750     0.7783   0.225    0.822
ldose         0.9060     0.1671   5.422 5.89e-08 ***
sexM:ldose    0.3529     0.2700   1.307    0.191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   4.9937  on  8  degrees of freedom
AIC: 43.104
```

Logistic Regression: Tolerance Distribution

```
> summary(budworm.lg <- glm(SF ~ sex*ldose, family = binomial))
```

Here, `sex*ldose` expands to `1 + sex + ldose + sex:ldose`

Thus, it specifies sex-specific submodels of the form:

If `sex=female`: $\eta = \beta_0 + \beta_{\text{ldose}}\texttt{ldose}$
If `sex=male`: $\eta = \left(\beta_0 + \beta_{\text{sexM}}\right) + \left(\beta_{\text{ldose}} + \beta_{\text{sexM:ldose}}\right)\texttt{ldose}$

Therefore, this interaction term in the model additionally allows for **sex-specific slopes**.

Logistic Regression: Tolerance Distribution

Alternative model specification by numerical vector with elements s_i/m_i, where m_i is the number of trials and s_i the number of successes. The values m_i are specified using `weights`.

```
> summary(glm(dead/20 ~ sex*ldose, family = binomial,
+             weights=rep(20,12)))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9935     0.5527  -5.416 6.09e-08 ***
sexM          0.1750     0.7783   0.225    0.822
ldose         0.9060     0.1671   5.422 5.89e-08 ***
sexM:ldose    0.3529     0.2700   1.307    0.191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Regression: Tolerance Distribution

Result indicates a **significant slope** of `ldose` for females.

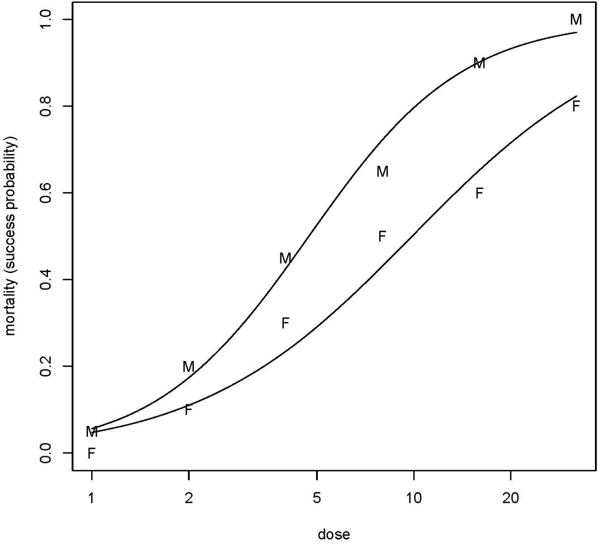
`sexM:ldose` represents (not significant) a larger slope for males.

First level of `sex` relates to female moths (`"F"` before `"M"`) described by the intercept.

`sexM` is the (not significant) difference of the sex-specific intercepts.

```
> plot(c(1,32), c(0,1), type="n", xlab="dose", log="x")
> text(2^ldose, dead/20, as.character(sex))
> ld <- seq(0, 5, 0.1), l <- length(ld)
> lines(2^ld, predict(budworm.lg, data.frame(ldose=ld,
+   sex=factor(rep("M",l,levels=levels(sex))),type="response"))
> lines(2^ld, predict(budworm.lg, data.frame(ldose=ld,
+   sex=factor(rep("F",l,levels=levels(sex))),type="response"))
```

Logistic Regression: Tolerance Distribution



Logistic Regression: Tolerance Distribution

`sexM` describes the difference at dose 1μg ($\log_2(\text{Dose}) = 0$) and seems to be irrelevant.

If we are interested in difference at dose 8μg ($\log_2(\text{Dose}) = 3$), we get

```
> summary(budworm.lg8 <- update(budworm.lg, .~sex*I(ldose-3)))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.2754     0.2305  -1.195  0.23215
sexM                1.2337     0.3770   3.273  0.00107 **
I(ldose - 3)        0.9060     0.1671   5.422 5.89e-08 ***
sexM:I(ldose - 3)   0.3529     0.2700   1.307  0.19117
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Regression: Tolerance Distribution

```
> anova(budworm.lg, test = "Chisq")

           Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                          11    124.876
sex         1    6.077        10    118.799   0.0137 *
ldose       1  112.042         9      6.757   <2e-16 ***
sex:ldose   1    1.763         8      4.994   0.1842
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Significant sex-difference at dose 8μg.

Model fits nicely (deviance 5 at df $= 8$).

Confirmed by the analysis of deviance.

We resign interactions.

Logistic Regression: Tolerance Distribution

Quadratic `ldose` term not necessary.

```
> anova(update(budworm.lg, .~.+ sex*I(ldose^2)), test="Chisq")

               Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                              11    124.876
sex             1    6.077        10    118.799   0.0137 *
ldose           1  112.042         9      6.757  <2e-16 ***
I(ldose^2)      1    0.907         8      5.851   0.3410
sex:ldose       1    1.240         7      4.611   0.2655
sex:I(ldose^2)  1    1.439         6      3.172   0.2303
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Analysis recommends a model with 2 parallel lines on the predictor- (logit)-axis (1 for each `sex`).

Logistic Regression: Tolerance Distribution

Estimate dose that guarantees a certain mortality: first reparameterize model, such that each sex has its own intercept.

```
> summary(budworm.lg0<-glm(SF~sex+ldose-1, family=binomial))

Coefficients:
      Estimate Std. Error z value Pr(>|z|)
sexF   -3.4732     0.4685  -7.413 1.23e-13 ***
sexM   -2.3724     0.3855  -6.154 7.56e-10 ***
ldose   1.0642     0.1311   8.119 4.70e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 126.2269  on 12  degrees of freedom
Residual deviance:   6.7571  on  9  degrees of freedom
AIC: 42.867
```

Logistic Regression: Tolerance Distribution

ξ_p is the value of $\log_2(\text{dose})$ inducing mortality p.

$2^{\xi_{0.5}}$ is the **50% lethal dose** (**LD50**) and using a link $g(p) = \beta_0 + \beta_1 \xi_p$ we get

$$\xi_p = \frac{g(p) - \beta_0}{\beta_1} \ .$$

Dose ξ_p depends on $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$, thus $\xi_p = \xi_p(\boldsymbol{\beta})$.

Replace $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ yields estimator $\hat{\xi}_p = \xi_p(\hat{\boldsymbol{\beta}})$ with property (linear approximation)

$$\hat{\xi}_p \approx \xi_p + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \frac{\partial \xi_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \ .$$

Because $\mathrm{E}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{\beta}$, we have $\mathrm{E}(\hat{\xi}_p) \approx \xi_p$.

Logistic Regression: Tolerance Distribution

Moreover, the delta method gives

$$\mathrm{var}(\hat{\xi}_p) = \frac{\partial \xi_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}\, \mathrm{var}(\hat{\boldsymbol{\beta}}) \frac{\partial \xi_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\,,$$

where

$$\frac{\partial \xi_p}{\partial \beta_0} = -\frac{1}{\beta_1}\,, \qquad \frac{\partial \xi_p}{\partial \beta_1} = -\frac{g(p) - \beta_0}{\beta_1^2} = -\frac{\xi_p}{\beta_1}\,.$$

Function `dose.p` from `MASS` gives for **female** moths:

```
> require(MASS)
> dose.p(budworm.lg0, cf = c(1,3), p = (1:3)/4) # females
           Dose     SE
p = 0.25: 2.231 0.2499
p = 0.50: 3.264 0.2298
p = 0.75: 4.296 0.2747
```

Logistic Regression: Tolerance Distribution

For **male** moths we get:

```
> dose.p(budworm.lg0, cf = c(2,3), p = (1:3)/4) # males
             Dose      SE
p = 0.25: 1.197 0.2635
p = 0.50: 2.229 0.2260
p = 0.75: 3.262 0.2550
```

An estimated dose of $\log_2(\text{dose}) = 3.264$, or $\text{dose} = 9.60$, is necessary to kill 50% of the female moths, but only $\text{dose} = 4.69$ for 50% of the male moths.

Logistic Regression: Tolerance Distribution

Alternative probit model: gives very similar results.

E.g., for **female** moths we get

```
> dose.p(update(budworm.lg0, family=binomial(link=probit)),
+        cf=c(1,3), p=(1:3)/4)
             Dose      SE
p = 0.25: 2.191 0.2384
p = 0.50: 3.258 0.2241
p = 0.75: 4.324 0.2669
```

Logistic Regression: Parameter Interpretation

Assume that the mean of a binary response depends on a two-level factor $x \in \{0, 1\}$.

Cell probabilities:

	$x = 1$	$x = 0$
$y = 1$	π_1	π_0
$y = 0$	$1 - \pi_1$	$1 - \pi_0$

For $x = 1$, the **odds** that $y = 1$ occurs and not $y = 0$ is

$$\pi_1/(1 - \pi_1).$$

Its log-transformation

$$\log \frac{\pi_1}{1 - \pi_1} = \text{logit}(\pi_1)$$

is called **log-odds** or **Logit**.

Logistic Regression: Parameter Interpretation

The ratio of the odds for $x = 1$ and the one for $x = 0$ is called **odds-ratio**

$$\psi = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} ,$$

Its log-transformation is the **log-odds ratio** or the **logit difference**

$$\log\psi = \log\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \mathrm{logit}(\pi_1) - \mathrm{logit}(\pi_0) .$$

Logistic Regression: Parameter Interpretation

Let $\mu(x) = \Pr(y = 1|x)$ and $1 - \mu(x) = \Pr(y = 0|x)$, $x \in \{0, 1\}$.
The model

$$\log \frac{\mu(x)}{1 - \mu(x)} = \beta_0 + \beta_1 x$$

gives probabilities

	$x = 1$	$x = 0$
$y = 1$	$\mu(1) = \dfrac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$	$\mu(0) = \dfrac{\exp(\beta_0)}{1 + \exp(\beta_0)}$
$y = 0$	$1 - \mu(1) = \dfrac{1}{1 + \exp(\beta_0 + \beta_1)}$	$1 - \mu(0) = \dfrac{1}{1 + \exp(\beta_0)}$

As log-odds ratio we get

$$\log \psi = \log \frac{\mu(1)/(1 - \mu(1))}{\mu(0)/(1 - \mu(0))} = \log \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \beta_1 \, .$$

Logistic Regression: Parameter Interpretation

For a general predictor x with a respective model, the odds are

$$\frac{\Pr(y=1|x)}{\Pr(y=0|x)} = \frac{\mu(x)}{1-\mu(x)} = \exp(\beta_0 + \beta_1 x) = \exp(\beta_0)\exp(\beta_1)^x \, .$$

Interpretation: for a unit change in x, the odds of $y = 1$ multiply by $\exp(\beta_1)$.

Logistic Regression: Parameter Interpretation

Remission Example: Injection treatment of 27 cancer patients should decay the carcinoma. The response measures whether a patient achieved remission.

Most important explanatory variable LI (labeling index) describes the cell activity after treatment.

For $n = 14$ different LI values, the response $m_i y_i$ is the number of successful remissions at m_i patients all with labeling index $\texttt{LI}_i$:

$\texttt{LI}_i$	m_i	$m_i y_i$	$\texttt{LI}_i$	m_i	$m_i y_i$	$\texttt{LI}_i$	m_i	$m_i y_i$
8	2	0	18	1	1	28	1	1
10	2	0	20	3	2	32	1	0
12	3	0	22	2	1	34	1	1
14	3	0	24	1	0	38	3	2
16	3	0	26	1	1			

Logistic Regression: Parameter Interpretation

Assumption: m_i patients in the LI_i group are homogenous, i.e.

$$m_i y_i \overset{ind}{\sim} \text{Binomial}(m_i, \mu_i)\,, \qquad \text{with} \qquad \log \frac{\mu_i}{1-\mu_i} = \beta_0 + \beta_1 \mathrm{LI}_i\,.$$

```
> li <- c(seq(8, 28, 2), 32, 34, 38)
> total <-c(2, 2, 3, 3, 3, 1, 3, 2, 1, 1, 1, 1, 1, 3)
> back  <-c(0, 0, 0, 0, 0, 1, 2, 1, 0, 1, 1, 0, 1, 2)
> SF <- cbind(back, nonback = total - back)
> summary(carcinoma <- glm(SF ~ li, family=binomial))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7771     1.3786   -2.74   0.0061 **
li            0.1449     0.0593    2.44   0.0146 *
---
    Null deviance: 23.961  on 13  degrees of freedom
Residual deviance: 15.662  on 12  degrees of freedom
AIC: 24.29
```

Logistic Regression: Parameter Interpretation

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7771     1.3786   -2.74   0.0061 **
li            0.1449     0.0593    2.44   0.0146 *
---
```

Interpretation:

- If LI increases by 1 unit, the odds for remission multiplies with $\exp(0.145) = 1.156$ (increases by 15.6%).
- Remission prob. is $1/2$ if $\hat{\eta} = 0$, i.e. if $\texttt{LI} = -\hat{\beta}_0/\hat{\beta}_1 = 26.07$.
- At the mean LI-value, $\sum_i \texttt{LI}_i m_i / \sum_i m_i = 20.07$, the linear predictor is $\hat{\beta}_0 + \hat{\beta}_1 20.07 = -0.8691$ (corresponds with 29.54%). There are 9 successes from 27 patients observed, i.e. 33.33%.

Logistic Regression: Parameter Interpretation

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7771     1.3786   -2.74   0.0061 **
li            0.1449     0.0593    2.44   0.0146 *
---
```

Interpretation:

• Logistic regression curve: $\mu(\eta) = e^{\eta}/(1 + e^{\eta})$ thus $\partial\mu(x)/\partial x = \beta_1\mu(x)(1 - \mu(x))$. Largest ascent in $\mu(x) = 1/2$, i.e. in $\mathtt{LI} = 26.07$, which is $\hat{\beta}_1/4 = 0.0362$.

• Question: does remission significantly depend on the `LI`-value? The p-value of 1.46% (Wald test) shows evidence for this.

Logistic Regression: Parameter Interpretation

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7771     1.3786   -2.74   0.0061 **
li            0.1449     0.0593    2.44   0.0146 *
---
    Null deviance: 23.961  on 13  degrees of freedom
Residual deviance: 15.662  on 12  degrees of freedom
AIC: 24.29
```

Interpretation:

- For an iid random sample model the (NULL) Deviance is 23.96 with $df = 13$. The deviance difference is 8.30 with associated loss of $df = 1$ corresponds to $\chi^2_{1;1-\alpha}$ quantile with $\alpha = 0.004$ (even more significant as Wald test).

Significant (positive) association between LI and remission.

Logistic Regression: Parameter Interpretation

Simpler with :

```
> anova(carcinoma, test="Chisq")

     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    13      23.96
li    1    8.299        12      15.66  0.00397 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Regression: Parameter Interpretation

Model with each patient remission as **Bernoulli** variable yields the same coefficients, but different values for the deviance and the degrees of freedom.

```
> index <- rep.int(li, times=total)
> B<-c(0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,1,0,0,1,1,0,1,1,1,0)
> summary(carcinomaB <- glm(B ~ index, family=binomial))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7771     1.3786   -2.74   0.0061 **
index         0.1449     0.0593    2.44   0.0146 *
---
    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 26.073  on 25  degrees of freedom
AIC: 30.07
```

Logistic Regression: Parameter Interpretation

Again, the deviance difference is the same as before:

```
> anova(carcinomaB, test="Chisq")

      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     26      34.37
index  1    8.299        25      26.07  0.00397 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Notice that the probability of remission ($y = 1$) is modeled again.

Because all $m_i = 1$ in case of Bernoullis, we do not need to explicitly specify `weights`.

Poisson Regression: Counts

Binomial responses: relative or absolute **frequencies**.

Poisson responses: **counts**.

Assumption: mean equals variance, i.e. $\mathrm{E}(y_i) = \mu_i = \mathrm{var}(y_i)$.

Is the Poisson probability function a member of the linear exponential family (LEF)?

Poisson Regression: Counts

$y \sim \text{Poisson}(\mu)$, $y = 0, 1, 2, \ldots,$ mean $\mu > 0$:

$$f(y|\mu) = \frac{\mu^y}{y!} e^{-\mu} = \exp\left(y \log \mu - \mu - \log y!\right).$$

Let $\theta = \log \mu$ and $\phi = 1$, then this is a member of the LEF with (weight $a = 1$)

$$b(\theta) = \exp(\theta), \qquad c(y, \phi) = -\log y!.$$

Canonical link is the **log-link**. **Dispersion** is **known** ($\phi = 1$). Moreover,

$$\begin{aligned} \mathrm{E}(y) &= b'(\theta) = \exp(\theta) = \mu \\ \mathrm{var}(y) &= b''(\theta) = \exp(\theta) = \mu. \end{aligned}$$

Poisson Regression: Counts

Log-linear model for counts:

$$y_i \overset{ind}{\sim} \text{Poisson}(\mu_i) \qquad \text{with} \quad \log(\mu_i) = \eta_i \, .$$

The (scaled) deviance equals ($\phi = 1$)

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \left\{ y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right\} \, .$$

If the model contains an intercept, this deviance simplifies to

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} y_i \log \frac{y_i}{\hat{\mu}_i} \, .$$

Deviance contribution is zero for $y_i = 0$ (independent of $\hat{\mu}_i$).

Poisson Regression: Counts

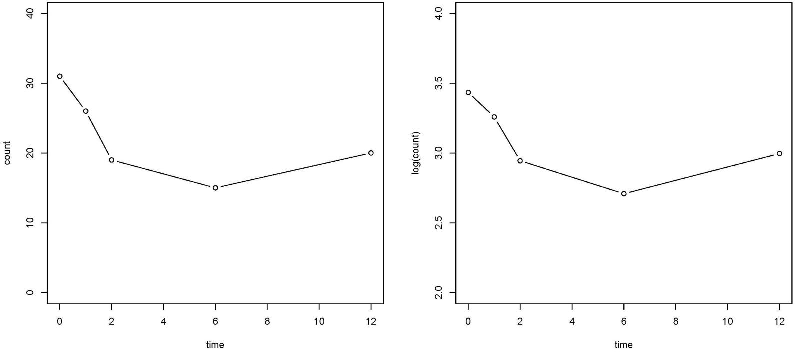
Example: Storing microorganisms (deep-frozen -70^{o}C). Bacterial concentration (counts in a fixed area) measured at initial freezing and then at 1, 2, 6, and 12 months afterwards.

time	0	1	2	6	12
count	31	26	19	15	20

Aim: model from which fractional recovery rates at specified times after freezing can be predicted.
Guess: some sort of exponential decay curve.

```
> time  <- c( 0, 1, 2, 6,12)
> count <- c(31,26,19,15,20)

> plot(time, count, type="b", ylim=c(0, 40))
> plot(time, log(count), type="b", ylim=c(2, 4))
```

Poisson Regression: Counts



Poisson Regression: Counts

We have expected exponential decay (but last observation is even larger than the two before).

Probably some measurement error causes this behavior.

Possibly log(concentration) depends linearly on time?

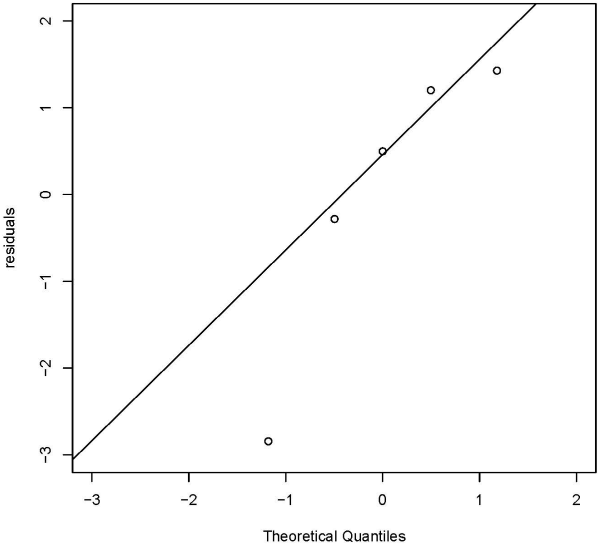
Test, if observed curvature is relevant, by allowing the quadratic term `time`^{2} in the model.

First assumption, counts follow a normal distribution and satisfy a linear model in `time` and `time`^{2}.

Poisson Regression: Counts

```
> summary(mo.lm <- lm(count ~ time + I(time^2)))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.80042    1.88294  15.827  0.00397 **
time        -4.61601    1.00878  -4.576  0.04459 *
I(time^2)    0.31856    0.08049   3.958  0.05832 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.438 on 2 degrees of freedom
Multiple R-squared:  0.9252,    Adjusted R-squared:  0.8503
F-statistic: 12.36 on 2 and 2 DF,  p-value: 0.07483

> qqnorm(residuals(mo.lm), ylab="residuals", xlim=c(-3,2),
+       ylim=c(-3,2), main="")
> qqline(residuals(mo.lm))
```

Poisson Regression: Counts



Poisson Regression: Counts

Quadratic term seems relevant (p-value 0.058).

Q-Q Plot: points deviate from straight line
$\Rightarrow$ normal assumptions seems unrealistic.

$\Rightarrow$ try Poisson model.

Usually Poisson-means are modeled on log-scale .

Is quadratic time effect still necessary in the model?

Poisson Regression: Counts

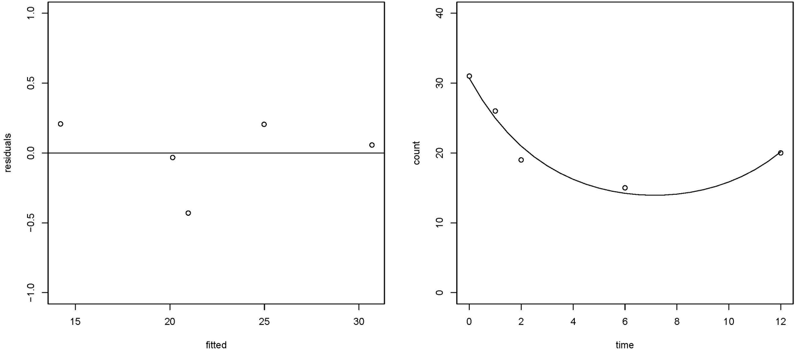
```
> summary(mo.P0 <- glm(count ~ time+I(time^2), family=poisson))
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.423818   0.149027  22.975   <2e-16 ***
time        -0.221389   0.095623  -2.315   0.0206 *
I(time^2)    0.015527   0.007731   2.008   0.0446 *
---
    Null deviance: 7.0672  on 4  degrees of freedom
Residual deviance: 0.2793  on 2  degrees of freedom
AIC: 30.849

> r <- residuals(mo.P0, type="pearson"); sum(r^2)
[1] 0.2745424
```

Under true model, deviance (0.2793) and $X^2 = 0.2745$ should correspond to about $df = n - p = 2$ (test on goodness–of–fit). Since both values are small, this does not argue against the Poisson assumption ($\text{var}(y_i) = \mu_i$).

Poisson Regression: Counts

```
> f <- fitted(mo.P0)
> plot(f, r, ylab="residuals", xlab="fitted", ylim=c(-1,1))
> abline(0,0)

> plot(time, count, ylim=c(0,40))
> time.new <- seq(0, 12, 0.5)
> lines(time.new, predict(mo.P0, data.frame(time=time.new),
+                             type="response"))
```

Poisson Regression: Counts



Poisson Regression: Counts

Residual plot: if variances equal means, the Pearson residual is

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

If we replace $\hat{\mu}_i$ with μ_i, then r_i should reflect mean zero and variance one.

Residual plot is relatively ($n = 5$) unremarkable. Poisson assumption seems applicable.

To validate the model quality (exploratively), we plot observed and fitted values against time. Of course, such a 3 parameter model has to fit well the 5 observations.

Poisson Regression: Counts

Measurement errors can also result in growing counts (but this is impossible in reality).

The Wald statistic indicated that `time`^{2} seems to be significant in the predictor (p-value 0.0446).

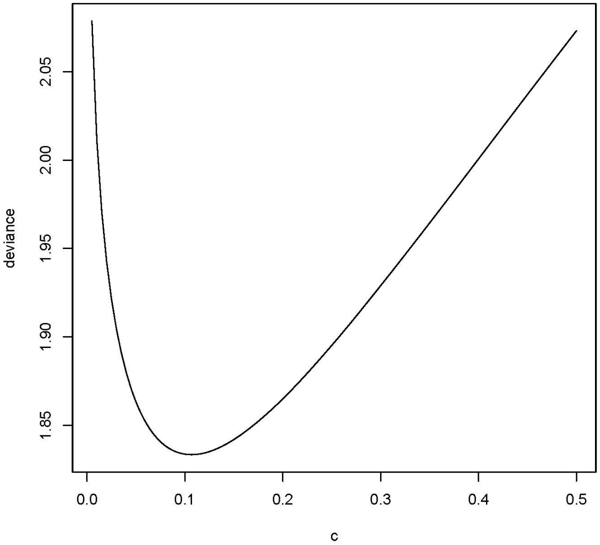
Possibly we get a more realistic model using log(`time`) instead of `time`.

Poisson Regression: Counts

If time has a multiplicative effect ($\mu \propto \texttt{time}^{\gamma}$), then the model should be based on $\log(\texttt{time})$ as predictor.
But then the starting time $\log(0)$ is problematic.
Therefore we consider the transformation $\log(\texttt{time} + c)$ with unknown positive shift c.
To determine c, we minimize the deviance in c, i.e.

```
> c <- d <- 1:100
> for (i in 1:100) {
+    c[i] <- i/200
+    d[i] <- deviance(glm(count ~ log(time+c[i]),
+                         family=poisson))
+    }
> plot(c, d, type="l", ylab="deviance")
> c[d==min(d)]
[1] 0.105
```

Poisson Regression: Counts



Poisson Regression: Counts

Optimal value of c under model $1 + \log(\texttt{time} + c)$ is $c = 0.105$ and $\log(\texttt{time} + 0.105)$ will be used from now on as predictor.

```
> time.c <- time + 0.105
> summary(mo.P3 <- glm(count ~ log(time.c), family=poisson))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.15110    0.09565  32.945   <2e-16 ***
log(time.c) -0.12751    0.05493  -2.321   0.0203 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7.0672  on 4  degrees of freedom
Residual deviance: 1.8335  on 3  degrees of freedom
AIC: 30.403
```

Poisson Regression: Counts

It is again advisable to consider also a model with quadratic time effect in order to check if there is still some curvature left.

```
> mo.P2 <- glm(count ~ log(time.c)+I(log(time.c)^2),
+               family=poisson)
> anova(mo.P3, mo.P2, test="Chisq")
Analysis of Deviance Table

Model 1: count ~ log(time.c)
Model 2: count ~ log(time.c) + I(log(time.c)^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         3     1.8335
2         2     1.7925  1  0.04109   0.8394
```

Quadratic effect is no longer necessary. It seems that when using the log-transformed shifted time, this linear effect suffices in the predictor.

Poisson Regression: Counts

Wanted: approximative pointwise CIV for $\mu_0 = \exp(\eta_0)$.

Idea 1: use $\hat{\eta}_0 = \mathbf{x}_0^\top \hat{\boldsymbol{\beta}}$ with $\widehat{s.e.}(\hat{\eta}_0)$. The transformed 95% interval is

$$CIV(\mu_0) = \Big(\exp\left(\hat{\eta}_0 \pm 1.96 \times \widehat{s.e.}(\hat{\eta}_0) \right) \Big) .$$

Idea 2: Delta method yields

$$\log \hat{\mu} \approx \log \mu + (\hat{\mu} - \mu) \frac{\partial \log \mu}{\partial \mu} ,$$

giving approximative variance, resp. standard error

$$\begin{aligned} \operatorname{var}(\log \hat{\mu}) &\approx \operatorname{var}(\hat{\mu}) \frac{1}{\mu^2} \\ \widehat{\operatorname{var}}(\hat{\mu}) &\approx \hat{\mu}^2 \operatorname{var}(\hat{\eta}) \qquad \Rightarrow \qquad \widehat{s.e.}(\hat{\mu}_0) \approx \hat{\mu}_0 \widehat{s.e.}(\hat{\eta}_0) . \end{aligned}$$

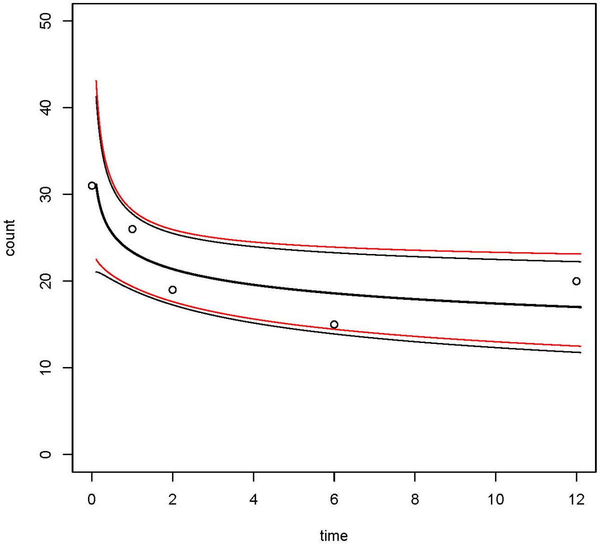
As 95% CIV we get

$$CIV_\Delta(\mu_0) = \Big(\hat{\mu}_0 \pm 1.96 \times \hat{\mu}_0 \widehat{s.e.}(\hat{\eta}_0) \Big) .$$

Poisson Regression: Counts

```
> # Delta-Method
> t.new <- data.frame(time.c = seq(0,12,.005) + 0.105)
> r.pred<-predict(mo.P3,newdata=t.new,type="response",se.fit=T)
> fit   <- r.pred$fit
> upper <- fit + qnorm(0.975)*r.pred$se.fit
> lower <- fit - qnorm(0.975)*r.pred$se.fit
> plot(time, count, type="p", xlab="time", ylab="count")
> lines(time.c.new[,1], upper)
> lines(time.c.new[,1], fit)
> lines(time.c.new[,1], lower)

> # using prediction of type="link"
> l.pred <- predict(mo.P3, newdata=t.new, type="link", se.fit=T)
> fit   <- exp(l.pred$fit)
> upper <- exp(l.pred$fit + qnorm(0.975)*l.pred$se.fit)
> lower <- exp(l.pred$fit - qnorm(0.975)*l.pred$se.fit)
> lines(time.c.new[,1], upper, col=2)
> lines(time.c.new[,1], lower, col=2)
```

Poisson Regression: Counts



Poisson Regression: Contingency Tables

Log-linear models to analyze if 2 factors are **stochastically independent**.
None of the 2 factors will be defined as response – we call them both **classificators**.

Example: Habitat of Lizards: counts on how many lizards have chosen what kind of perch, characterized by two-level factors: `height` (≥ 4.75, < 4.75) and `diameter` (≤ 4.0, > 4.0). The following counts have been observed:

Perch		`diameter`		
		≤ 4.0	> 4.0	total
	≥ 4.75	61	41	102
`height`	< 4.75	73	70	143
total		134	111	245

Poisson Regression: Contingency Tables

Question: are `diameter` and `height` classifications independent? Association is measurable by **odds-ratios**. In case of independence, the odds-ratio is 1. We get as estimate

$$\hat{\psi} = \frac{61/41}{73/70} = \frac{61/73}{41/70} = 1.43\,.$$

Does this indicate that for the true parameter $\psi \neq 1$ holds?

We introduce a log-linear model for 2×2 tables and define the following observed counts:

	B		
A	1	2	total
1	y_{11}	y_{12}	$y_{1\bullet}$
2	y_{21}	y_{21}	$y_{2\bullet}$
total	$y_{\bullet 1}$	$y_{\bullet 2}$	$y_{\bullet\bullet}$

with $y_{\bullet\bullet} = n$, the sample size.

Poisson Regression: Contingency Tables

If y_{kl} are Poisson counts and we use a log-link function and A and B as explanatory predictors, this would correspond to a log-linear model.

Distributions of A and of B (marginals) are not of interest.

We consider the next two models

1. $A + B$ (independence),
2. $A * B \equiv A + B + A : B$ (dependence, saturated model).

Poisson Regression: Contingency Tables

Independence Model:

Assumption: for all pairs (a_i, b_i), $i = 1, \ldots, n$, the probability to fall in cell (k, l) is π_{kl}. Then

$$\mathrm{E}(y_{kl}) = \mu_{kl} = n \cdot \pi_{kl} \,, \qquad k, l \in \{1, 2\} \,.$$

In case of stochastic independence, i.e. if

$$\pi_{kl} = \Pr(A = k, B = l) = \Pr(A = k)\Pr(B = l) = \pi_k^A \pi_l^B \,,$$

then the associated log-linear model is

$$\log \mu_{kl} = \log n + \log \pi_k^A + \log \pi_l^B \,.$$

The logarithm of the expected count in cell (k, l) is an additive function of the k-th row effect and the l-th column effect. Thus

$$\log \mu_{kl} = \lambda + \lambda_k^A + \lambda_l^B \,, \qquad k, l \in \{1, 2\} \,.$$

Poisson Regression: Contingency Tables

$$\log \mu_{kl} = \lambda + \lambda_k^A + \lambda_l^B \,, \qquad k, l \in \{1,2\} \,.$$

How to define the parameters, and how many are identifiable?

If a contrast parametrization is of interest, we define

$$\lambda_k^A = \log \pi_k^A - \frac{1}{2}\sum_{h=1}^{2} \log \pi_h^A$$

$$\lambda_l^B = \log \pi_l^B - \frac{1}{2}\sum_{h=1}^{2} \log \pi_h^B$$

$$\lambda = \log n + \frac{1}{2}\sum_{h=1}^{2} \log \pi_h^A + \frac{1}{2}\sum_{h=1}^{2} \log \pi_h^B \,.$$

With this parametrization (deviation from the means) we have

$$\sum_{k=1}^{2} \lambda_k^A = \sum_{k=1}^{2} \left\{ \log \pi_k^A - \frac{1}{2}\sum_{h=1}^{2} \log \pi_h^A \right\} = 0 = \sum_{l=1}^{2} \lambda_l^B \,.$$

Poisson Regression: Contingency Tables

$$\sum_{k=1}^{2} \lambda_k^A = \sum_{k=1}^{2} \left\{ \log \pi_k^A - \frac{1}{2} \sum_{h=1}^{2} \log \pi_h^A \right\} = 0 = \sum_{l=1}^{2} \lambda_l^B .$$

Besides λ there is only 1 row and 1 column parameter identifiable. For both others $\lambda_2^A = -\lambda_1^A$, $\lambda_2^B = -\lambda_1^B$ hold.

This model is called log-linear **independence model**.
The respective predictors are

	B	
A	1	2
1	$\lambda + \lambda_1^A + \lambda_1^B$	$\lambda + \lambda_1^A - \lambda_1^B$
2	$\lambda - \lambda_1^A + \lambda_1^B$	$\lambda - \lambda_1^A - \lambda_1^B$

Poisson Regression: Contingency Tables

Alternative parametrization: **reference cell** instead of contrasts. Characterize an arbitrary cell as reference and define parameters, that describe the deviations from this reference cell.
If e.g. cell $(1,1)$ is the reference, this gives

$$\begin{aligned}\lambda_k^A &= \log \pi_k^A - \log \pi_1^A \\ \lambda_l^B &= \log \pi_l^B - \log \pi_1^B \\ \lambda &= \log n + \log \pi_1^A + \log \pi_1^B\end{aligned}$$

with identifiability constraints

$$\lambda_1^A = \lambda_1^B = 0\,.$$

The respective predictors are

	B	
A	1	2
1	λ	$\lambda + \lambda_2^B$
2	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_2^B$

Poisson Regression: Contingency Tables

Notice that this (reference cell) parametrization results in

$$\begin{aligned}\log\psi &= \log\frac{\mu_{11}/\mu_{12}}{\mu_{21}/\mu_{22}}\\ &= \log\mu_{11} - \log\mu_{12} - \log\mu_{21} + \log\mu_{22}\\ &= \lambda - (\lambda + \lambda_2^B) - (\lambda + \lambda_2^A) + (\lambda + \lambda_2^A + \lambda_2^B)\\ &= 0\,.\end{aligned}$$

Thus, an odds-ratio of $\psi = 1$ is equivalent with independence.

This holds independently of the choice of the reference cell.

Poisson Regression: Contingency Tables

Saturated (full) Model:
If no independence can be assumed we define

$$\log \mu_{kl} = \lambda + \lambda_k^A + \lambda_l^B + \lambda_{kl}^{AB}\,, \qquad k, l \in \{1, 2\}\,.$$

The interaction parameters λ_{kl}^{AB} describe the discrepancies from the independence model.
If contrasts should be used, then the parameters are based on the linear predictors $\eta_{kl} = \log \mu_{kl}$.Let

$$\eta_{k\bullet} = \frac{1}{2}\sum_{l=1}^{2} \eta_{kl}\,, \quad \eta_{\bullet l} = \frac{1}{2}\sum_{k=1}^{2} \eta_{kl}\,, \quad \eta_{\bullet\bullet} = \lambda = \frac{1}{2}\frac{1}{2}\sum_{k=1}^{2}\sum_{l=1}^{2} \eta_{kl}\,.$$

Poisson Regression: Contingency Tables

Define row effects λ_k^A, column effects λ_l^B, and interaction effects λ_{kl}^{AB} as deviations from the mean predictor

$$\begin{aligned}
\lambda_k^A &= \eta_{k\bullet} - \eta_{\bullet\bullet} \\
\lambda_l^B &= \eta_{\bullet l} - \eta_{\bullet\bullet} \\
\lambda_{kl}^{AB} &= \eta_{kl} - \eta_{k\bullet} - \eta_{\bullet l} + \eta_{\bullet\bullet} = \underbrace{(\eta_{kl} - \eta_{\bullet\bullet})}_{\eta_{kl}-\lambda} - \underbrace{(\eta_{k\bullet} - \eta_{\bullet\bullet})}_{\lambda_k^A} - \underbrace{(\eta_{\bullet l} - \eta_{\bullet\bullet})}_{\lambda_l^B} .
\end{aligned}$$

λ_k^A, λ_l^B denote deviations from the predictor mean λ.
λ_{kl}^{AB} are cell effects that are adjusted for row and column effects.
Since all parameters are centered around their means we have

$$\sum_{k=1}^{2} \lambda_k^A = \sum_{l=1}^{2} \lambda_l^B = 0 .$$

Thus, again only 1 free row and 1 free column parameter.

Poisson Regression: Contingency Tables

For the interactions we get

$$\begin{aligned}\sum_{k=1}^{2} \lambda_{kl}^{AB} &= \sum_{k=1}^{2} \eta_{kl} - \sum_{k=1}^{2} \eta_{k\bullet} - 2\eta_{\bullet l} + 2\eta_{\bullet\bullet} \\ &= 2\eta_{\bullet l} - 2\eta_{\bullet\bullet} - 2\eta_{\bullet l} + 2\eta_{\bullet\bullet} = 0 = \sum_{l=1}^{2} \lambda_{kl}^{AB} \; .\end{aligned}$$

Because of this, the sum of all interactions in each row and in each column is 0.

In case of a 2 × 2 table there is only 1 free interaction parameter!

Poisson Regression: Contingency Tables

The independence model is a special case of the full model with $\lambda_{kl}^{AB} = 0$ for all (k, l).

The additional parameters λ_{kl}^{AB} are **association parameters**, describing the deviations from independence between A and B.

The total number of free parameters is 3 under the independence model and 4 in case of the dependence model.

Default approach in R is to use a `treatment` parametrization, i.e. a reference cell $(1, 1)$. If a `sum` parametrization should be used, then (for *unordered* and *ordered* factors)

```
> options(contrasts=c("contr.sum", "contr.poly"))
```

We can change back to the `treatment` parametrization through

```
> options(contrasts=c("contr.treatment", "contr.poly"))
```

Poisson Regression: Contingency Tables

It's again simpler to work with a reference cell, e.g. cell $(1, 1)$. Setting $\lambda = \eta_{11}$ gives

$$\begin{aligned} \lambda_k^A &= \eta_{k1} - \eta_{11} \\ \lambda_l^B &= \eta_{1l} - \eta_{11} \\ \lambda_{kl}^{AB} &= \eta_{kl} - \eta_{k1} - \eta_{1l} + \eta_{11} = \underbrace{(\eta_{kl} - \eta_{11})}_{\eta_{kl} - \lambda} - \underbrace{(\eta_{k1} - \eta_{11})}_{\lambda_k^A} - \underbrace{(\eta_{1l} - \eta_{11})}_{\lambda_l^B} . \end{aligned}$$

Thus $\lambda_1^A = \lambda_1^B = 0$. Moreover all interactions in the first row and in the first column are 0 and we get

	B	
A	1	2
1	λ	$\lambda + \lambda_2^B$
2	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_2^B + \lambda_{22}^{AB}$

Poisson Regression: Contingency Tables

What are the MLEs of these parameters?

$$\log \hat{\mu}_{11} = \hat{\lambda} = \log y_{11}$$

$$\log \hat{\mu}_{21} = \hat{\lambda} + \hat{\lambda}_2^A = \log y_{21} \Rightarrow \hat{\lambda}_2^A = \log y_{21} - \log y_{11} = \log \frac{y_{21}}{y_{11}}$$

$$\log \hat{\mu}_{12} = \hat{\lambda} + \hat{\lambda}_2^B = \log y_{12} \Rightarrow \hat{\lambda}_2^B = \log y_{12} - \log y_{11} = \log \frac{y_{12}}{y_{11}}$$

$$\log \hat{\mu}_{22} = \hat{\lambda} + \hat{\lambda}_2^A + \hat{\lambda}_2^B + \hat{\lambda}_{22}^{AB} = \log y_{22}$$
$$\Rightarrow \hat{\lambda}_{22}^{AB} = \log y_{22} - \log y_{11} - \log \frac{y_{21}}{y_{11}} - \log \frac{y_{12}}{y_{11}} = \log \frac{y_{11}y_{22}}{y_{12}y_{21}} .$$

MLE of the interaction effect is the observed log-odds-ratio, that estimates the deviation from the independence model.

Poisson Regression: Contingency Tables

Example: Habitat of Lizards

To use cell $(1, 1)$ as reference in R, we need e.g.

```
> count <- c(61, 41, 73, 70)

> (hei <-   factor(c(">4.75", ">4.75", "<4.75", "<4.75")))
[1] >4.75 >4.75 <4.75 <4.75
Levels: <4.75 >4.75

> (height <- relevel(hei, ref = ">4.75"))
[1] >4.75 >4.75 <4.75 <4.75
Levels: >4.75 <4.75

> diameter <- factor(c("<4.0", ">4.0", "<4.0", ">4.0"))
```

Poisson Regression: Contingency Tables

```
> summary(dep<-glm(count ~ height * diameter, family=poisson))

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)               4.1109     0.1280  32.107   <2e-16 ***
height<4.75               0.1796     0.1735   1.035   0.3006
diameter>4.0             -0.3973     0.2019  -1.967   0.0491 *
height<4.75:diameter>4.0  0.3553     0.2622   1.355   0.1754
---

    Null deviance:  1.0904e+01  on 3  degrees of freedom
Residual deviance: -8.8818e-16  on 0  degrees of freedom
AIC: 31.726
```

Poisson Regression: Contingency Tables

Deviance = 0 on *df* = 0. Model reproduces the data exactly. Estimated odds-ratio is

```
> exp(dep$coef[4])
height<4.75:diameter>4.0
                1.426662
```

Under the independence model we get

```
> summary(ind<-glm(count ~ height + diameter, family=poisson))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.0216     0.1148  35.023  < 2e-16 ***
height<4.75    0.3379     0.1296   2.607  0.00913 **
diameter>4.0  -0.1883     0.1283  -1.467  0.14231
---
    Null deviance: 10.9036  on 3  degrees of freedom
Residual deviance:  1.8477  on 1  degrees of freedom
AIC: 31.574
```

Poisson Regression: Contingency Tables

Odds-ratio is 0 now and the deviance increases by 1.85. This can be used as test statistic on $H_0 : \psi = 1$ giving a p-value of

```
> pchisq(ind$deviance, 1, lower.tail = FALSE)
[1] 0.174055
```

Evidence for a non-significant improvement (compare with p-value 0.1754 of the respective Wald statistic). Thus we cannot reject $H_0 : \psi = 1$ and `diameter` and `height` seem to classify independently!

Poisson Regression: Contingency Tables

More than two-level factors:
Results can be generalized for **multi-level** classifying factors. Let A be a K-level and B a L-level factor. The **independence model** is

$$\log \mu_{kl} = \lambda + \lambda_k^A + \lambda_l^B \,, \qquad k = 1, \ldots, K, \; l = 1, \ldots, L \,.$$

With cell $(1,1)$ as reference we define

$$\begin{aligned} \lambda_k^A &= \log \pi_k^A - \log \pi_1^A \\ \lambda_l^B &= \log \pi_l^B - \log \pi_1^B \\ \lambda &= \log n + \log \pi_1^A + \log \pi_1^B \end{aligned}$$

and the same set of identifiability conditions hold, i.e.

$$\lambda_1^A = \lambda_1^B = 0 \,.$$

There are $1 + (K-1) + (L-1)$ parameter freely estimable.

Poisson Regression: Contingency Tables

Respective predictors are

	B					
A	1	2	$\cdots$	l	$\cdots$	L
1	λ	$\lambda + \lambda_2^B$	$\cdots$	$\lambda + \lambda_l^B$	$\cdots$	$\lambda + \lambda_L^B$
2	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_2^B$	$\cdots$	$\lambda + \lambda_2^A + \lambda_l^B$	$\cdots$	$\lambda + \lambda_2^A + \lambda_L^B$
$\vdots$						
k	$\lambda + \lambda_k^A$	$\lambda + \lambda_k^A + \lambda_2^B$	$\cdots$	$\lambda + \lambda_k^A + \lambda_l^B$	$\cdots$	$\lambda + \lambda_k^A + \lambda_L^B$
$\vdots$						
K	$\lambda + \lambda_K^A$	$\lambda + \lambda_K^A + \lambda_2^B$	$\cdots$	$\lambda + \lambda_K^A + \lambda_l^B$	$\cdots$	$\lambda + \lambda_K^A + \lambda_L^B$

Poisson Regression: Contingency Tables

MLEs are now for $k = 1, \dots, K$ and $l = 1, \dots, L$

$$\log \hat{\mu}_{11} = \hat{\lambda} = \log \frac{y_{1\bullet} y_{\bullet 1}}{y_{\bullet\bullet}}$$

$$\log \hat{\mu}_{k1} = \hat{\lambda} + \hat{\lambda}_k^A = \log \frac{y_{k\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} \Rightarrow \hat{\lambda}_k^A = \log \frac{y_{k\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} - \log \frac{y_{1\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} = \log \frac{y_{k\bullet}}{y_{1\bullet}}$$

$$\log \hat{\mu}_{1l} = \hat{\lambda} + \hat{\lambda}_l^B = \log \frac{y_{1\bullet} y_{\bullet l}}{y_{\bullet\bullet}} \Rightarrow \hat{\lambda}_l^B = \log \frac{y_{1\bullet} y_{\bullet l}}{y_{\bullet\bullet}} - \log \frac{y_{1\bullet} y_{\bullet 1}}{y_{\bullet\bullet}} = \log \frac{y_{\bullet l}}{y_{\bullet 1}}$$

Poisson Regression: Contingency Tables

The **saturated model** for a $K \times L$ table is

$$\log \mu_{kl} = \lambda + \lambda_k^A + \lambda_l^B + \lambda_{kl}^{AB}\,, \qquad k = 1, \ldots, K,\ \ l = 1, \ldots, L\,.$$

With reference cell $(1, 1)$ we get for all $k = 1, \ldots, K$, $l = 1, \ldots, L$

$$\begin{aligned} \lambda_k^A &= \eta_{k1} - \eta_{11} \\ \lambda_l^B &= \eta_{1l} - \eta_{11} \\ \lambda_{kl}^{AB} &= \eta_{kl} - \eta_{k1} - \eta_{1l} + \eta_{11} = \underbrace{(\eta_{kl} - \eta_{11})}_{\eta_{kl} - \lambda} - \underbrace{(\eta_{k1} - \eta_{11})}_{\lambda_k^A} - \underbrace{(\eta_{1l} - \eta_{11})}_{\lambda_l^B}, \end{aligned}$$

where $\lambda_1^A = \lambda_1^B = 0$.
Again, all interactions in row 1 and in column 1 are 0.
Thus, the total number of estimable parameters is
$1 + (K - 1) + (L - 1) + (K - 1)(L - 1) = K \times L$.

Poisson Regression: Contingency Tables

The predictors are defined as:

				B		
A	1	2	$\cdots$	l	$\cdots$	L
1	λ	$\lambda + \lambda_2^B$	$\cdots$	$\lambda + \lambda_l^B$	$\cdots$	$\lambda + \lambda_L^B$
2	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_2^B + \lambda_{22}^{AB}$	$\cdots$	$\lambda + \lambda_2^A + \lambda_l^B + \lambda_{2l}^{AB}$	$\cdots$	$\lambda + \lambda_2^A + \lambda_L^B + \lambda_{2L}^{AB}$
$\vdots$						
k	$\lambda + \lambda_k^A$	$\lambda + \lambda_k^A + \lambda_2^B + \lambda_{k2}^{AB}$	$\cdots$	$\lambda + \lambda_k^A + \lambda_l^B + \lambda_{kl}^{AB}$	$\cdots$	$\lambda + \lambda_k^A + \lambda_L^B + \lambda_{kL}^{AB}$
$\vdots$						
K	$\lambda + \lambda_K^A$	$\lambda + \lambda_K^A + \lambda_2^B + \lambda_{K2}^{AB}$	$\cdots$	$\lambda + \lambda_K^A + \lambda_l^B + \lambda_{Kl}^{AB}$	$\cdots$	$\lambda + \lambda_K^A + \lambda_L^B + \lambda_{KL}^{AB}$

Saturated model allows for $(K-1)(L-1)$ additional parameters than the independence model.

Poisson Regression: Contingency Tables

Example: Recurrences of Cervical Cancer
Are the predictive factors border zone (BZ) involvement and number affected lymph node (LN) stations classifying independently?
Consider the following counts:

	LN stations			
	0	1	2	≥ 3
BZ not involved	124	21	16	13
BZ involved	58	12	7	5
more than BZ inv.	14	19	12	12

We first fit the saturated model to the data and then test on necessary interactions.

Poisson Regression: Contingency Tables

```
> anova(glm(total ~ B*L, family=poisson), test="Chisq")
Analysis of Deviance Table
     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    11    316.184
B     2   69.569         9    246.615 7.821e-16 ***
L     3  203.594         6     43.021 < 2.2e-16 ***
B:L   6   43.021         0      0.000 1.155e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There is evidence, that the 6 interaction parameter are unequal 0 and thus the independence hypothesis can be rejected.

Poisson Regression: Contingency Tables

Alternatively, we consider the Pearson statistic under the independence model, i.e.

$$X^2 = \sum_{i=1}^{3} \sum_{j=1}^{4} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

with $\log \mu_{ij} = \lambda + \lambda_i^B + \lambda_j^L$. Its realization is

```
> ind <- glm(total ~ B+L, family=poisson)
> r <- residuals(ind, type="pearson")
> sum(r^2)
[1] 43.83645
```

and equals the χ^2 test statistic in the analysis of contingency tables.

Poisson Regression: Contingency Tables

Pearson statistic can be also directly calculated as

```
> (N <- matrix(total, 3, 4, byrow=TRUE))
     [,1] [,2] [,3] [,4]
[1,]  124   21   16   13
[2,]   58   12    7    5
[3,]   14   19   12   12
> chisq.test(N)

        Pearson's Chi-squared test

data:  N
X-squared = 43.8365, df = 6, p-value = 7.965e-08
```