Basic Methods of Data Analysis
Part 3

Sepp Hochreiter

Institute of Bioinformatics
Johannes Kepler University, Linz, Austria

Chapter 5

Linear Models

Linear Models

We have considered linear regression for bivariate variables:

$$y = a + b\,x$$

dependent variable: y

explanatory variable, independent variable, regressor: x

parameters: a and b

Goal: fitting a linear function (a line) to data points
Objective: sum of the squared deviations between data and the line (regression line, least squares line)

multivariate case:

generalize x to a vector of features $\boldsymbol{x}$ with components x_j which are called explanatory variables, independent variables, regressors, features

→ multiple linear regression: vector $\boldsymbol{x}$ (considered here)

→ multivariate linear regression: vector y

Linear Models

Linear Regression

The Linear Model

m features $x_1, \ldots, x_m$ summarized by the vector $\boldsymbol{x} = (x_1, \ldots, x_m)$

linear model:

$$y = \beta_0 + \sum_{j=1}^{m} x_j \beta_j + \epsilon$$

$(m + 1)$ parameters $\beta_0, \beta_1, \ldots, \beta_m$

additive noise or error: ϵ

simplify notation: $\boldsymbol{x} = (1, x_1, \ldots, x_m)$ $\quad$ $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)$

$$y = \boldsymbol{x}^T \boldsymbol{\beta} + \epsilon$$

Linear Models

n observations $\{(y_i, \boldsymbol{x}_i) \mid 1 \leq i \leq n\}$

y_i summarized by y

$\boldsymbol{x}_i$ summarized by $\boldsymbol{X} \in \mathbb{R}^{n \times (m+1)}$

$\boldsymbol{\epsilon}_i$ summarized by $\boldsymbol{\epsilon} \in \mathbb{R}^n$

matrix equation:

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

Interpretations

explanatory / independent variables:

- random variables sampled together with the dependent variable
- constants which are fixed

Linear Models

Regression models

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

β_0 is y-intercept and β_1 is the slope

example with 7 observations

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \\ 1 & x_6 \\ 1 & x_7 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{pmatrix}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Two regressors:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \\ 1 & x_{51} & x_{52} \\ 1 & x_{61} & x_{62} \\ 1 & x_{71} & x_{72} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{pmatrix}$$

Linear Models

one-way ANOVA or cell means model

Example: study design has 3 groups → find the mean for each group

$$y_{gi} = \beta_g + \epsilon_{gi}$$

β_g is the mean of group g

Example: 3 examples for first group, two examples for other groups

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}$$

Linear Models

another example of an ANOVA model

study design: control group / reference group vs. treatment groups

control is group 1: offset of group g from group 1 is β_g $\Rightarrow \beta_1 = 0$

$$y_{gi} = \beta_0 + \beta_g + \epsilon_{gi}$$

Example: 3 in group 1, 2 in group 2, and 2 in group 3

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}$$

- mean of the reference group 1 is β_0
- difference to the reference group is β_g
- per design: $\beta_1 = 0$

Linear Models

two-way ANOVA model

two known groupings or two known factors

Each observation belongs
- to a group of the first grouping and
- to a group of the second grouping

$$y_{ghi} = \beta_0 + \beta_g + \alpha_h + (\beta\alpha)_{gh} + \epsilon_{ghi}$$

g denotes the first factor (grouping) and h the second
i indicates the replicate for this combination of factors

main effects: β_g and α_h
interaction effects between the factors: $(\beta\alpha)_{gh}$

No unique solution if each combination of groups is observed once

⇒ noise free observations allow more than one set of parameters

→ avoid over-parametrization: additional constraints

Linear Models

constraints are either

- **sum-to-zero constraint**: main and interaction effect parameters sum to zero for each index or
- **corner point parametrization**: all parameters that contain the index 1 are zero

corner point parametrization:

$$\begin{aligned} \alpha_1 &= 0 \\ \beta_1 &= 0 \\ (\beta\alpha)_{1h} &= 0 \\ (\beta\alpha)_{g1} &= 0 \end{aligned}$$

Example: 1. factor (β_g) 3 levels, 2. factor (α_h) 2 levels, two replicates

$$\begin{pmatrix} y_{111} \\ y_{112} \\ y_{211} \\ y_{212} \\ y_{311} \\ y_{312} \\ y_{121} \\ y_{122} \\ y_{221} \\ y_{222} \\ y_{321} \\ y_{322} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_3 \\ \alpha_2 \\ (\beta\alpha)_{22} \\ (\beta\alpha)_{32} \end{pmatrix} + \begin{pmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{311} \\ \epsilon_{312} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{221} \\ \epsilon_{222} \\ \epsilon_{321} \\ \epsilon_{322} \end{pmatrix}$$

Linear Models

Assumptions:

- **strict exogeneity**: zero mean error given regressors $\mathrm{E}(\boldsymbol{\epsilon} \mid \boldsymbol{X}) = \mathbf{0}$
  $\Rightarrow \mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$
  $\Rightarrow \mathrm{E}(\boldsymbol{X}^T \boldsymbol{\epsilon}) = \mathbf{0}$

- **linear independence**: $\Pr(\mathrm{rank}(\boldsymbol{X}) = m + 1) = 1$
  otherwise estimation is only possible in the data subspace

- (desired) second moments should be finite: to ensure $\mathrm{E}(\frac{1}{n} \boldsymbol{X}^T \boldsymbol{X})$
  $\Rightarrow$ to derive theoretical properties

- **spherical errors** : $\mathrm{Var}(\boldsymbol{\epsilon} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$

  $\Rightarrow$ homoscedasticity: error has the same variance in each observation $\mathrm{E}(\epsilon_i^2 \mid \boldsymbol{X}) = \sigma^2$ otherwise: weighted least squared

  $\Rightarrow$ errors not correlated $\mathrm{E}(\epsilon_i \epsilon_k \mid \boldsymbol{X}) = 0$ for $i \neq k$

Linear Models

Normality of the Errors given the regressors:

$$\boldsymbol{\epsilon} \mid \boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$$

⇒ least squared estimator is the maximum likelihood estimator
⇒ asymptotically efficient (asymptotically the best estimator)
⇒ test hypotheses based on the normality assumption

Samples $\{(y_i, \boldsymbol{x}_i)\}$ independent and identically distributed (iid)

$$\Pr((y_i, \boldsymbol{x}_i) \mid (y_1, \boldsymbol{x}_1), \ldots, (y_{i-1}, \boldsymbol{x}_{i-1}), (y_{i+1}, \boldsymbol{x}_{i+1}), \ldots, (y_n, \boldsymbol{x}_n)) = \Pr((y_i, \boldsymbol{x}_i))$$

$$\Pr((y_i, \boldsymbol{x}_i)) = \Pr((y_k, \boldsymbol{x}_k))$$

iid exogeneity: $\mathrm{E}(\epsilon_i \mid \boldsymbol{x}_i) = 0$

iid linear independence: $\mathrm{Var}(\boldsymbol{x}) = \mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(\boldsymbol{x}_i \boldsymbol{x}_i^T) = \mathrm{E}(\frac{1}{n} \boldsymbol{X}^T \boldsymbol{X})$

iid homoscedasticity: $\mathrm{Var}(\epsilon_i \mid \boldsymbol{x}_i) = \sigma^2$

Linear Models

time series models: iid assumption does not hold

New assumptions:

- **stationary** stochastic process (probability distribution shift invariant)
- **ergodic** stochastic process (time average is the population average)
- **predetermined** regressors: $\mathrm{E}(\boldsymbol{x}_i \epsilon_i) = 0 \text{ for all } i = 1, \ldots, n$
- **full rank** matrix: $\mathrm{rank}(\mathrm{E}(\boldsymbol{x}_i \boldsymbol{x}_i^T)) = m + 1$
- **martingale difference sequence** (zero mean given the past) of $\{\boldsymbol{x}_i \epsilon_i\}$ with existing second moments $\mathrm{E}(\epsilon_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T)$

Linear models for time series: **autoregressive models**

Least Squares Parameter Estimation

residual of i-th observation: $r_i = y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}$

$\tilde{\boldsymbol{\beta}}$: candidate for the parameter vector β

sum of squared residuals (SSR), **error sum of squares** (ESS), or **residual sum of squares** (RSS):

$$S(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}\right)^2 = \left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{y} - \boldsymbol{X}\,\tilde{\boldsymbol{\beta}}\right)$$

least squares estimator $\hat{\boldsymbol{\beta}}$ minimizes $S(\tilde{\boldsymbol{\beta}})$:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\tilde{\boldsymbol{\beta}}} S(\tilde{\boldsymbol{\beta}}) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

The solution is obtained by setting the derivative to zero:

$$\frac{\partial S(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = 2\,\boldsymbol{X}^T\left(\boldsymbol{y} - \boldsymbol{X}\,\tilde{\boldsymbol{\beta}}\right) = \boldsymbol{0}$$

pseudo inverse of $\boldsymbol{X}$: $\boldsymbol{X}^+ = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$

$$\boldsymbol{X}^+\boldsymbol{X} = \boldsymbol{I}_m$$

Linear Models

least squares estimator:

- minimal variance linear unbiased estimator (MVLUE) → best linear unbiased estimator
- normality assumption for the errors → maximum likelihood estimator (MLE) → asymptotically efficient

β true parameter vector

$\tilde{\beta}$ candidate parameter vector (variable)

$\hat{\beta}$ estimator (optimal value of and objective)

Evaluation and Interpretation of the Estimation

Residuals and Error Variance

estimated values: $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\,\boldsymbol{X}^{+}\boldsymbol{y} = \boldsymbol{X}\,\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}$

hat matrix (projection matrix): $\boldsymbol{P} = \boldsymbol{X}\,\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$

"puts a hat on y"

$\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{P}^2 = \boldsymbol{P}$

minimal residuals: $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I}_n - \boldsymbol{P})\,\boldsymbol{y} = (\boldsymbol{I}_n - \boldsymbol{P})\,\boldsymbol{\epsilon}$

$\boldsymbol{P}$ and $(\boldsymbol{I}_n - \boldsymbol{P})$ are symmetric and idempotent

$\sigma^2 \approx S(\hat{\boldsymbol{\beta}})$ is an estimate for the error:

$$S(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{y}^T\boldsymbol{y} - 2\,\hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}$$

$$= \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} = \hat{\boldsymbol{\epsilon}}^T\boldsymbol{y}$$

$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}$

Linear Models

least squares estimate (unbiased) for σ^2 : $s^2 = \frac{1}{n-m-1} S(\hat{\boldsymbol{\beta}})$

maximum likelihood estimate (biased) for σ^2: $\hat{\sigma}^2 = \frac{1}{n} S(\hat{\boldsymbol{\beta}})$

minimal mean squared estimate for σ^2: $\tilde{\sigma}^2 = \frac{1}{n-m+1} S(\hat{\boldsymbol{\beta}})$

covariance of residuals:

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T) &= (\boldsymbol{I}_n - \boldsymbol{P})\ \mathrm{E}(\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}^T)\ (\boldsymbol{I}_n - \boldsymbol{P}) \\ &= \sigma^2 (\boldsymbol{I}_n - \boldsymbol{P})^2 = \sigma^2 (\boldsymbol{I}_n - \boldsymbol{P}) \end{aligned}$$

covariance structure of error according to assumptions: $\sigma^2 \boldsymbol{I}_n$

Linear Models

coefficient of determination R^2

ratio of the variance "explained" by the model to the "total'" y-variance

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} = \frac{\boldsymbol{y}^T \boldsymbol{P}^T \boldsymbol{L}\, \boldsymbol{P}\, \boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{L}\, \boldsymbol{y}} = 1 - \frac{\boldsymbol{y}^T (\boldsymbol{I} - \boldsymbol{P})\, \boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{L}\, \boldsymbol{y}} = 1 - \frac{\text{SSR}}{\text{TSS}}$$

$\boldsymbol{L} = \boldsymbol{I}_n - (1/n)\mathbf{1}\,\mathbf{1}^T$

$\mathbf{1}$ is n-dimensional vector of ones

"TSS": sum of squares

"SSR": sum of squared residuals denoted by S

R^2 is between 0 and 1

the closer R^2 is to 1, the better the fit

constant offset (intercept) → X contains a column of ones

Outliers and Influential Observations

outlier: worse fitted than other observations → large error

influential observation: large effect on the model fitting / inferences

Outliers need not be influential observations

standardized or studentized residuals ρ_i: $$\rho_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - P_{ii}}}$$

P_{ii} are the diagonal elements of the hat matrix
$\hat{\sigma}^2$ is an estimate of the error variance

ρ_i can be used to check the fitted model and the model assumptions

Linear Models

Alternative is leave-one-out regression: one observation is removed from the data set and a least squares estimate performed

least squares estimator $\hat{\boldsymbol{\beta}}_{(i)}$ on the data where $(y_i, \boldsymbol{x}_i)$ is left out:

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{\hat{\epsilon}_i}{1 - P_{ii}} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i$$

residual of the left-out observation: $\hat{\epsilon}_{(i)} = \frac{\hat{\epsilon}_i}{1 - P_{ii}}$

Outlier detection:

- plotting the leave-one-out residuals against the standard residuals may reveal outliers
- Small $(1 - P_{ii})$ (hat matrix diagnal entries close to one)

Linear Models

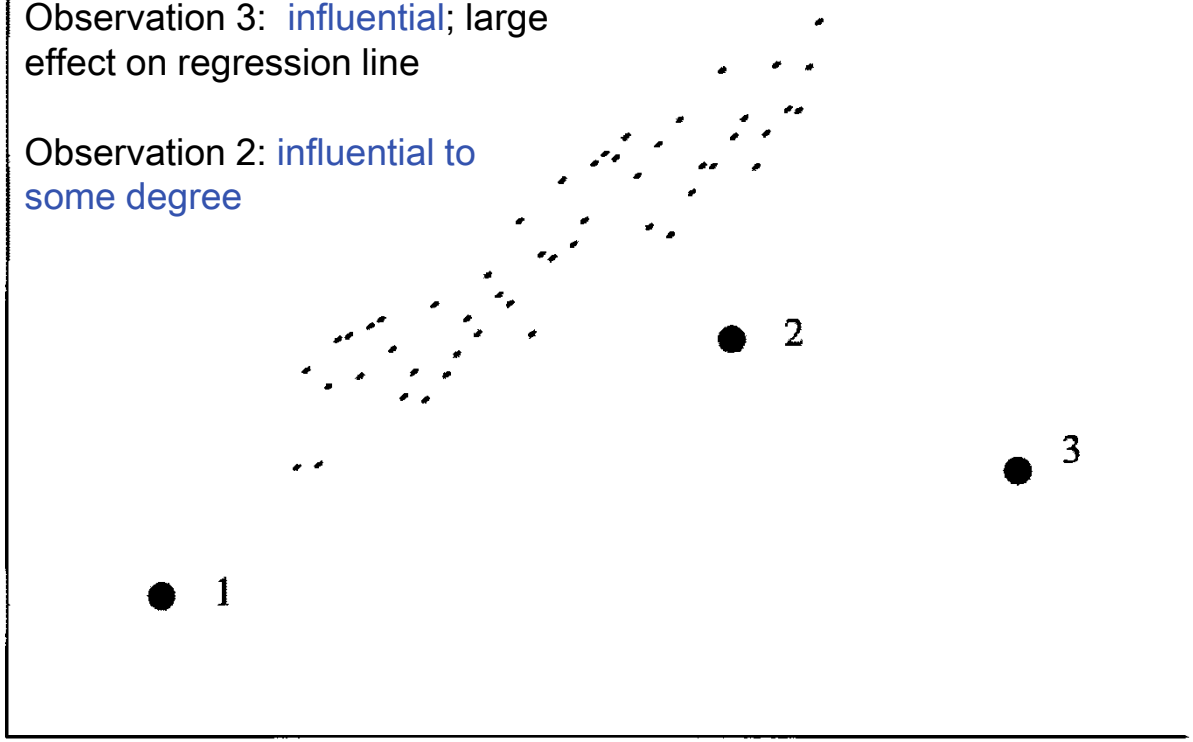
influential observation: large effect on parameter or output estimates

⇒ different estimates if influential observations is removed

Observation 1: not influential; close to the regression line

Observation 3: influential; large effect on regression line

Observation 2: influential to some degree



Linear Models

$$\hat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y} \qquad \hat{y}_i = \sum_{j=1}^{n} P_{ij} y_j = P_{ii} y_i + \sum_{j, j \neq i} P_{ij} y_j$$

P_{ii} large → P_{ij} for $j \neq i$ small ($\boldsymbol{P}$ idempotent)

Leverage of y_i : P_{ii} (y_i's contribution to its estimate)

Cook's distance (influence of the i-th observation): can be measured

$$D_i = \frac{\rho_i^2}{m+1} \frac{P_{ii}}{1 - P_{ii}}$$

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(m+1)\, s^2} = \frac{(\boldsymbol{X}\, \hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{X}\, \hat{\boldsymbol{\beta}})^T (\boldsymbol{X}\, \hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{X}\, \hat{\boldsymbol{\beta}})}{(m+1)\, s^2} = \frac{(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})^T (\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})}{(m+1)\, s^2}$$

D_i is proportional to the Euclidean distance between the estimate using all data and the estimate where the i-th observation is removed

same result by leave-one-out estimate

Linear Models

Confidence Intervals for Parameters and Prediction

Normally Distributed Error Terms

error normally distributed → least squares is maximum likelihood:

$$\hat{\boldsymbol{\beta}} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\beta},\ \sigma^2\ (\boldsymbol{X}^T\ \boldsymbol{X})^{-1}\right) \qquad \xrightarrow{d}$$

convergence in distribution

maximum likelihood estimator is optimal for unbiased estimators

This distribution gives an approximated two-sided confidence interval:

$$\beta_j \in \left[\ \hat{\beta}_j \pm t_{\alpha/2,n-m-1}\ s\ \sqrt{\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]_{jj}}\ \right]$$

$t_{\alpha/2,n-m-1}$: upper $\alpha/2$ percentage point of the central t-distribution

α: desired significance level of the test → $100(1\text{-}\alpha)\%$ confident that the interval contains the true β_j

confidence intervals do not hold simultaneously for all β_j

Linear Models

confidence interval for the noise free prediction:

$$\boldsymbol{x}^T\boldsymbol{\beta} \in \left[\boldsymbol{x}^T\hat{\boldsymbol{\beta}} \pm t_{\alpha/2,n-m-1}\, s \sqrt{\boldsymbol{x}^T\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\,\boldsymbol{x}} \right]$$

holds for a single prediction but not for multiple predictions

With noise ϵ and $\hat{y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$

$$y \in \left[\hat{y} \pm t_{\alpha/2,n-m-1}\, s \sqrt{1 + \boldsymbol{x}^T\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\,\boldsymbol{x}} \right]$$

estimator s^2 is not efficient and chi-squared distributed:

$$s^2 \sim \frac{\sigma^2}{n-m-1}\,\chi^2_{n-m-1} \quad \text{with variance} \quad 2\sigma^4/(n-m-1)$$

estimator is the minimal variance unbiased estimator (MVUE)

estimator is independent of $\hat{\boldsymbol{\beta}}$ (advantageous for tests)

confidence interval for σ^2: $\frac{(n-m-1)\,s^2}{\chi^2_{\alpha/2,n-m-1}} \leq \sigma^2 \leq \frac{(n-m-1)\,s^2}{\chi^2_{1-\alpha/2,n-m-1}}$

Error Term Distribution Unknown

least squares estimator is consistent, that is, converges in probability to the true value → law of large number and central limit theorem
estimator is asymptotically normally distributed:

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\ \sigma^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

$$\hat{\boldsymbol{\beta}} \sim_a \mathcal{N}\left(\boldsymbol{\beta},\ \frac{\sigma^2}{n}\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

difference to normal errors

$\sim_a$ asymptotically distributed

approximated two-sided confidence interval:

$$\beta_j \in \left[\ \hat{\beta}_j \pm q^{\mathcal{N}(0,1)}_{1-\alpha/2}\ \sqrt{\frac{1}{n}\hat{\sigma}^2\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]_{jj}}\ \right]$$

q is the quantile function of the standard normal distribution
$(1\text{-}\alpha)$ confidence level

Linear Models

fourth moment of ϵ exists → LSE for σ^2 consistent and asympt. normal

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} \mathcal{N}\left(0,\ \mathrm{E}(\epsilon^4) - \sigma^4\right)$$

$$\hat{\sigma}^2 \sim_a \mathcal{N}\left(\sigma^2\ ,\ (\mathrm{E}(\epsilon^4) - \sigma^4) / n\right)$$

predicted response is a random variable:

$$\sqrt{n}\ (\hat{y} - y) \xrightarrow{d} \mathcal{N}\left(0,\ \sigma^2\ \boldsymbol{x}^T\ (\boldsymbol{X}^T \boldsymbol{X})^{-1}\ \boldsymbol{x}\right)$$

$$\hat{y} \sim_a \mathcal{N}\left(y,\ \frac{\sigma^2}{n}\ \boldsymbol{x}^T\ (\boldsymbol{X}^T \boldsymbol{X})^{-1}\ \boldsymbol{x}\right)$$

confidence interval for mean response (error bar on the prediction):

$$y \in \left[\ \boldsymbol{x}^T \hat{\boldsymbol{\beta}} \pm q_{1-\alpha/2}^{\mathcal{N}(0,1)} \sqrt{\tfrac{1}{n}\ \hat{\sigma}^2\ \boldsymbol{x}^T\ (\boldsymbol{X}^T \boldsymbol{X})^{-1}\ \boldsymbol{x}}\ \right]$$

(1-α) confidence level

Tests of Hypotheses

test whether independent variables are relevant for the regression
null hypothesis: without variables same fitting quality
null hypothesis rejected → variables relevant

Test for a Set of Variables Equal to Zero

remove h variables from the original model and fit a reduced model

$$\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2) \qquad \boldsymbol{X}_1 \in \mathbb{R}^{n \times (m-h+1)} \qquad \boldsymbol{X}_2 \in \mathbb{R}^{n \times h}$$

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \qquad \boldsymbol{\beta}_1 \in \mathbb{R}^{m-h+1} \text{ and } \boldsymbol{\beta}_2 \in \mathbb{R}^h$$

null hypothesis: $\boldsymbol{\beta}_2 = \mathbf{0}$

least squares estimator $\boldsymbol{X}_1$: $\hat{\boldsymbol{\beta}}_r \in \mathbb{R}^{m-h+1}$

$\hat{\boldsymbol{\beta}}_1 \in \mathbb{R}^{m-h+1}$
first (m-h+1) components of the least squares estimator

Linear Models

F statistic: $$F = \frac{\boldsymbol{y}^T \left(\boldsymbol{P} - \boldsymbol{P}_1\right) \boldsymbol{y} \;/\; h}{\boldsymbol{y}^T \left(\boldsymbol{I} - \boldsymbol{P}\right) \boldsymbol{y} \;/\; (n-m-1)} = \frac{\left(\hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}_r^T \boldsymbol{X}_1^T \boldsymbol{y}\right) \;/\; h}{\left(\boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y}\right) \;/\; (n-m-1)}$$

$\hat{\boldsymbol{\beta}}$: least squares estimator of the full model
$\hat{\boldsymbol{\beta}}_r$: least squares estimator of the reduced model

distribution of the F statistic:

- If H_0: $\boldsymbol{\beta}_2 = \boldsymbol{0}$ is false, F is distributed according to $F(h, n-m-1, \lambda)$

  $$\lambda = \boldsymbol{\beta}_2^T \left(\boldsymbol{X}_2^T \boldsymbol{X}_2 - \boldsymbol{X}_2^T \boldsymbol{X}_1 \left(\boldsymbol{X}_1^T \boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_2\right) \boldsymbol{\beta}_2 \;/\; (2\sigma^2)$$

- If H_0: $\boldsymbol{\beta}_2 = \boldsymbol{0}$ is true, λ=0 and F is distributed acc. to $F(h, n-m-1)$

H_0 is rejected if $F \geq F_{\alpha,h,n-m-1}$ (upper α percentage of the central F distribution)

F statistic expressed by R^2: $$F = \frac{\left(R^2 - R_1^2\right) \;/\; h}{\left(1 - R^2\right) \;/\; (n-m-1)}$$

R^2 : coefficient of determination for the full model
$R_1{}^2$: coefficient of determination for the reduced model

Linear Models

hypotheses tests summarized by Analysis-of-Variance (ANOVA) table:

Source of Variation	Degrees of freedom	Sum of squares	Mean square
reduced $\boldsymbol{\beta}_r$	$\mathrm{df} = m - h + 1$	$S = \hat{\boldsymbol{\beta}}_r^T \boldsymbol{X}_1^T \boldsymbol{y}$	$S \,/\, \mathrm{df}$
improved $\boldsymbol{\beta}$	$\mathrm{df} = h$	$S = \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}_r^T \boldsymbol{X}_1^T \boldsymbol{y}$	$S \,/\, \mathrm{df}$
residual	$\mathrm{df} = n - m - 1$	$S = \boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y}$	$S \,/\, \mathrm{df}$
total center	$\mathrm{df} = n - 1$	$S = \boldsymbol{y}^T \boldsymbol{y} - n\, \bar{y}$	$S \,/\, \mathrm{df}$
total	$\mathrm{df} = n$	$S = \boldsymbol{y}^T \boldsymbol{y}$	$S \,/\, \mathrm{df}$

ANOVA table for F test of H_0: $\boldsymbol{\beta}_2 = \boldsymbol{0}$

Linear Models

Test for a Single Variable Equal to Zero

F statistic of $H_0\ \beta_j = 0$: $F = \dfrac{\hat{\beta}_j^2}{s^2\ \left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]_{jj}}$

If $H_0\ \beta_j = 0$ is true, F is distributed according to $F(1, n-m-1)$

H_0 rejected if $F \geq F_{\alpha,1,(n-m-1)}$ (p-value smaller than α)

Alternatively, the t-statistic can be used: $t_j = \dfrac{\hat{\beta}_j}{s\ \sqrt{\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]_{jj}}}$

H_0 rejected if $|t_j| \geq t_{\alpha/2,(n-m-1)}$ (p-value smaller than α)

If several tests for parameters being zero are made:
correction for multiple testing → false discovery rate (FDR) or familywise α level can be adjusted

Hematology Data

six hematology variables measured on 51 workers:

- y: lymphocyte count,
- x_1: hemoglobin concentration,
- x_2: packed-cell volume,
- x_3: white blood cell count (times .01),
- x_4: neutrophil count,
- x_5: serum lead concentration.

Linear Models

#	y	x_1	x_2	x_3	x_4	x_5	#	y	x_1	x_2	x_3	x_4	x_5
1	14	13.4	39	41	25	17	27	16	15.5	45	52	30	20
2	15	14.6	46	50	30	20	28	18	14.5	43	39	18	25
3	19	13.5	42	45	21	18	29	17	14.4	45	60	37	23
4	23	15.0	46	46	16	18	30	23	14.6	44	47	21	27
5	17	14.6	44	51	31	19	31	43	15.3	45	79	23	23
6	20	14.0	44	49	24	19	32	17	14.9	45	34	15	24
7	21	16.4	49	43	17	18	33	23	15.8	47	60	32	21
8	16	14.8	44	44	26	29	34	31	14.4	44	77	39	23
9	27	15.2	46	41	13	27	35	11	14.7	46	37	23	23
10	34	15.5	48	84	42	36	36	25	14.8	43	52	19	22
11	26	15.2	47	56	27	22	37	30	15.4	45	60	25	18
12	28	16.9	50	51	17	23	38	32	16.2	50	81	38	18
13	24	14.8	44	47	20	23	39	17	15.0	45	49	26	24
14	26	16.2	45	56	25	19	40	22	15.1	47	60	33	16
15	23	14.7	43	40	13	17	41	20	16.0	46	46	22	22
16	9	14.7	42	34	22	13	42	20	15.3	48	55	23	23
17	18	16.5	45	54	32	17	43	20	14.5	41	62	36	21
18	28	15.4	45	69	36	24	44	26	14.2	41	49	20	20
19	17	15.1	45	46	29	17	45	40	15.0	45	72	25	25
20	14	14.2	46	42	25	28	46	22	14.2	46	58	31	22
21	8	15.9	46	52	34	16	47	61	14.9	45	84	17	17
22	25	16.0	47	47	14	18	48	12	16.2	48	31	15	18
23	37	17.4	50	86	39	17	49	20	14.5	45	40	18	20
24	20	14.3	43	55	31	19	50	35	16.4	49	69	22	24
25	15	14.8	44	42	24	29	51	38	14.7	44	78	34	16
26	9	14.9	43	43	32	17							

Linear Models

correlation matrix

```
cor(hemData)
           [,1]        [,2]      [,3]       [,4]       [,5]        [,6]
[1,] 1.00000000  0.23330745 0.2516182 0.79073232 0.02264257  0.08290783
[2,] 0.23330745  1.00000000 0.7737330 0.27650957 0.05537581 -0.08376682
[3,] 0.25161817  0.77373300 1.0000000 0.30847841 0.07642710  0.12970593
[4,] 0.79073232  0.27650957 0.3084784 1.00000000 0.60420947  0.07147757
[5,] 0.02264257  0.05537581 0.0764271 0.60420947 1.00000000  0.03169314
[6,] 0.08290783 -0.08376682 0.1297059 0.07147757 0.03169314  1.00000000
```

largest correlation between y and an explanatory variable is 0.79 (x_3)

Computing Estimates, Confidence Intervals, Tests

mean $\bar{y}$ of the response variable y:

```
(by <- mean(y))
[1] 22.98039
```

means of the explanatory variables:

```
(bx <- as.vector(colMeans(x)))
[1] 15.10784 45.19608 53.82353 25.62745 21.07843
```

covariance of the explanatory variables:

```
(Sxx <- var(x))
           [,1]      [,2]       [,3]       [,4]       [,5]
[1,]  0.6907373  1.494431   3.255412  0.3509804 -0.2966275
[2,]  1.4944314  5.400784  10.155294  1.3545098  1.2843137
[3,]  3.2554118 10.155294 200.668235 65.2729412  4.3141176
[4,]  0.3509804  1.354510  65.272941 58.1584314  1.0298039
[5,] -0.2966275  1.284314   4.314118  1.0298039 18.1537255
```

covariance between the response and the explanatory variables:

```
(syx <- as.vector(var(y,x)))
[1]   1.878157   5.663922 108.496471   1.672549   3.421569
```

Linear Models

Assuming centered data, we compute $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$

$$\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} = 1/n \left(\mathrm{Cov}(\boldsymbol{X})\right)^{-1}$$

$$\boldsymbol{X}^T \boldsymbol{y} = n \ \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{X})$$

$$\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \ = \ \left(\mathrm{Cov}(\boldsymbol{X})\right)^{-1} \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{X})$$

```
(bbeta <- solve(Sxx)%*%syx)
            [,1]
[1,] -0.21318219
[2,] -0.28884109
[3,]  0.85984756
[4,] -0.92921309
[5,]  0.05380269
```

Next we estimate β_0:

```
(bbeta0 <- by-t(syx)%*%solve(Sxx)%*%bx)
         [,1]
[1,] 15.65486
```

Linear Models

first column of X contains 1's to account for the intercept:

```
x1 <- cbind(rep(1,51),x)
(b1 <- solve(crossprod(x1))%*%t(x1)%*%y)
             [,1]
[1,] 15.65485611
[2,] -0.21318219
[3,] -0.28884109
[4,]  0.85984756
[5,] -0.92921309
[6,]  0.05380269
```

$$s^2 = \frac{1}{n-m-1} S(\hat{\boldsymbol{\beta}}) \qquad S(\hat{\boldsymbol{\beta}}) = \boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y}$$

n=51 and m=5

```
(s2 <- (crossprod(y)-t(b1)%*%t(x1)%*%y)/(51-6))
       [,1]
[1,] 4.3729
sqrt(s2)
         [,1]
[1,] 2.091148
```

Linear Models

coefficient of determination $R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$

```
fitted <- x1%*%b1
(R2 <- var(fitted)/var(y))
          [,1]
[1,] 0.9580513
```

coefficient of determination close to 1 → most data variance explained

approximate two-sided confidence intervals:

$$\beta_j \in \left[\hat{\beta}_j \pm t_{\alpha/2,n-m-1} \, s \sqrt{\left[(\boldsymbol{X}^T \boldsymbol{X})^{-1} \right]_{jj}} \right]$$

```
bup <- b1 - qt(0.025,45)*s*sqrt(diag(solve(crossprod(x1))))
blow <- b1 + qt(0.025,45)*s*sqrt(diag(solve(crossprod(x1))))
cbind(blow,bup)
             [,1]       [,2]
[1,]  3.03587336 28.2738389
[2,] -1.40187932  0.9755149
[3,] -0.71833021  0.1406480
[4,]  0.80366905  0.9160261
[5,] -1.02844916 -0.8299770
[6,] -0.09389755  0.2015029
```

Interval does not include zero

Estimated values

```
[1] 15.65485611
[2] -0.21318219
[3] -0.28884109
[4]  0.85984756
[5] -0.92921309
[6]  0.05380269
```

Linear Models

testing whether the components are significantly different from zero:

$$t_j = \frac{\hat{\beta}_j}{s\sqrt{\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]_{jj}}}$$

```
(t <- b1/(s*sqrt(diag(solve(crossprod(x1))))))
            [,1]
[1,]   2.4986561
[2,]  -0.3612114
[3,]  -1.3545299
[4,]  30.8271243
[5,] -18.8593854
[6,]   0.7336764
```

p-values:

```
2*pt(-abs(t),45)
              [,1]
[1,] 1.618559e-02
[2,] 7.196318e-01
[3,] 1.823298e-01
[4,] 6.694743e-32
[5,] 5.395732e-23
[6,] 4.669514e-01
```

only the intercept, x_3, and x_4 are significant, where the latter two are highly significant.

Assumption from intervals confirmed.

Using Predefined R Functions

```
l1 <- lm(y ~ x)
l1

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)           x1           x2           x3           x4           x5
    15.6549      -0.2132      -0.2888       0.8598      -0.9292       0.0538

anova(l1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          5 4494.2  898.84  205.55 < 2.2e-16 ***
Residuals 45  196.8    4.37
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear Models


```
summary(l1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6860 -0.9580  0.3767  1.0973  4.1742

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.65486    6.26531   2.499   0.0162 *
x1          -0.21318    0.59019  -0.361   0.7196
x2          -0.28884    0.21324  -1.355   0.1823
x3           0.85985    0.02789  30.827   <2e-16 ***
x4          -0.92921    0.04927 -18.859   <2e-16 ***
x5           0.05380    0.07333   0.734   0.4670
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.091 on 45 degrees of freedom
Multiple R-squared:  0.9581,    Adjusted R-squared:  0.9534
F-statistic: 205.5 on 5 and 45 DF,  p-value: < 2.2e-16
```

x_3 and x_4 are highly significant while intercept is significant

All values agree exactly with our computations.

Linear Models

confidence intervals:

```
confint(l1)
                  2.5 %     97.5 %
(Intercept)  3.03587336 28.2738389
x1          -1.40187932  0.9755149
x2          -0.71833021  0.1406480
x3           0.80366905  0.9160261
x4          -1.02844916 -0.8299770
x5          -0.09389755  0.2015029
```

again all values agree with those we have computed

Linear Models

AIC (Akaike information criterion) to compare models:

```
extractAIC(l1)
[1]  6.00000 80.86343

drop1(l1, test = "F")
Single term deletions

Model:
y ~ x
       Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>               196.8  80.863
x       5    4494.2 4691.0 232.600  205.55 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

drop1(l1, test = "Chisq")
Single term deletions

Model:
y ~ x
       Df Sum of Sq    RSS     AIC  Pr(>Chi)
<none>               196.8  80.863
x       5    4494.2 4691.0 232.600 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear Models

Carbohydrate Diet Data

twenty male insulin-dependent diabetics on high-carbohydrate diet for six months:

- response (carbohydrate),
- age,
- weight, and
- protein: percentages of total calories obtained from complex carbohydrates.

The response was thought to be related to other variables which are treated as explanatory variables.

Carbohydrate	Age	Weight	Protein
y	x_1	x_2	x_3
33	33	100	14
40	47	92	15
37	49	135	18
27	35	144	12
30	46	140	15
43	52	101	15
34	62	95	14
48	23	101	17
30	32	98	15
38	42	105	14
50	31	108	17
51	61	85	19
30	63	130	19
36	40	127	20
41	50	109	15
42	64	107	16
46	56	117	18
24	61	100	13
35	48	118	18
37	28	102	14

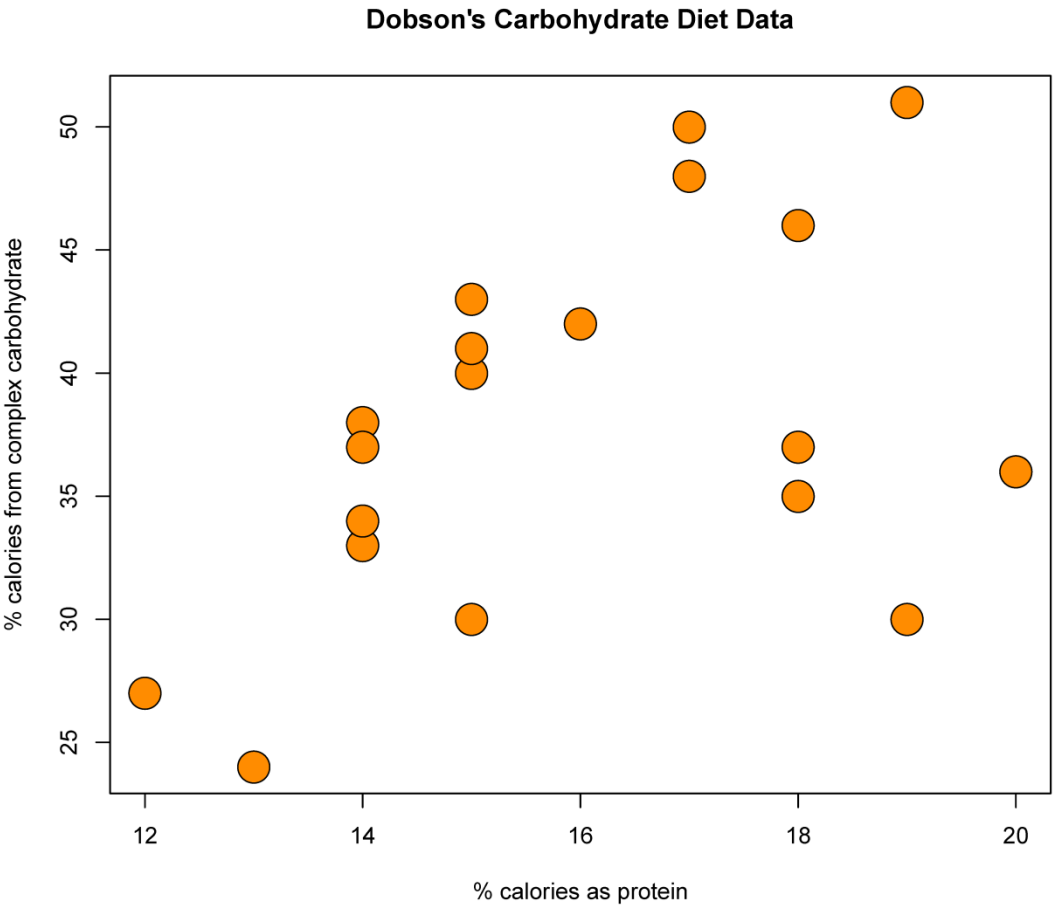
Linear Models

linear model fitted by least squares:

```
summary(lmcal <- lm(carb~age+wgt+prot, data= calorie))

Call:
lm(formula = carb ~ age + wgt + prot, data = calorie)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3424  -4.8203   0.9897   3.8553   7.9087

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.96006   13.07128   2.828  0.01213 *
age         -0.11368    0.10933  -1.040  0.31389
wgt         -0.22802    0.08329  -2.738  0.01460 *
prot         1.95771    0.63489   3.084  0.00712 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.956 on 16 degrees of freedom
Multiple R-squared:  0.4805,    Adjusted R-squared:  0.3831
F-statistic: 4.934 on 3 and 16 DF,  p-value: 0.01297
```

Linear Models

The feature Protein seems to be the feature that is most related to carbohydrates. We verify this by a scatter plot:



A linear dependence supports that Protein is related to carbohydrate

Analysis-of-variance (ANOVA)

linear models to compare means of responses to different treatments = levels of one factor

fitting linear models are analyzed by the variance explained

x neither measured nor a sample but constructed → dummy variables

X: design matrix (may not have full rank)

different groups corresponding to a factor

Linear Models

One Factor

response variable has two indices:

- group to which the observation belongs
- replicate number

The standard case is a treatment-control study, where one group are controls and the other group are the treatments. It is possible to analyze different treatments if they are mutually exclusive.

response variable: y_{gi}

$$y_{11}, y_{12}, \ldots, y_{1n_1}, y_{21}, y_{22}, \ldots, y_{2n_2}, y_{31}, \ldots, y_{Gn_G}$$

j-th group has n_j replicates and G denotes the number of groups

$$y_{gi} = \beta_0 + \beta_g + \epsilon_{gi}$$

β_0: constant offset or the mean of group 1 if β_1=0
β_g: mean difference to the offset (or group 1)
ϵ_{gi}: additive error term

Linear Models

Example: case-control study with 3 controls and 3 cases

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{pmatrix}$$

linear model: $\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$

X has lower rank than parameters (rank 2 because of identical rows)

→ least squares estimator is not computable: $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ does not exist

→ model is not identifiable (for data exists more than one solution)

Linear Models

Solutions:

i. **re-parametrization** using fewer parameters, e.g., corner point parametrization,
ii. **side conditions** as constraints on the parameters, e.g., sum-to-zero constraints,
iii. **linear projections** $\boldsymbol{a}^T\boldsymbol{\beta}$ of parameter vector.

ad (i) re-parametrization: $\beta_1 = 0$ corner point parametrization
→ offset of group g to the group 1 which are the controls

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{pmatrix}$$

corner point parametrization: remove all variables that contain the index one (group one is reference group).

Linear Models

General re-parametrization:

$$\boldsymbol{\gamma} = \boldsymbol{U}\ \boldsymbol{\beta}$$

$$\boldsymbol{X} = \boldsymbol{Z}\ \boldsymbol{U}$$

$$\boldsymbol{y} = \boldsymbol{Z}\ \boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

$\boldsymbol{Z}$ has full rank and $\boldsymbol{U}$ blows $\boldsymbol{Z}$ up to $\boldsymbol{X}$

Linear Models

ad (ii) side conditions: $\beta_1 + \beta_2 = 0 \rightarrow \beta_2 = -\beta_1$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{pmatrix}$$

β_2 is removed from these equations

sum-to-zero constraint: $\sum_{g=1}^{G} \beta_g = 0$

→ β_0 is the overall mean $\frac{1}{G} \sum_{g=0}^{G} \beta_g = \frac{1}{G} \beta_0$

estimate the deviation of the mean of a group from the overall mean

sum-to-zero constraints: sums over an index are set to zero

Linear Models

ad (iii) linear projection: for example $\boldsymbol{a} = (0, 1, -1))$

$$\beta_1' = \boldsymbol{a}^T \boldsymbol{\beta} = \beta_1 - \beta_2$$

specific questions can answered: difference of means of group 1 and 2

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{pmatrix} \beta_1' + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{pmatrix}$$

common null hypothesis is H_0: $\beta_1 = \beta_2 = \ldots = \beta_G$
means of all groups are equal

new variables $\beta_1^* = \beta_1 - \beta_2, \beta_2^* = \beta_1 - \beta_3, \ldots, \beta_{G-1}^* \beta_1 - \beta_G$

tested for $\beta_1^* = \beta_2^* = \ldots = \beta_{G-1}^* = 0$

Or constraint $\sum_{g=1}^{G} \beta_g = 0$

Two Factors

response variable has now three indices:

- group for the first factor
- group for the second factor
- replicate number

model $y_{ghi} = \beta_0 + \beta_g + \alpha_h + (\alpha\beta)_{gh} + \epsilon_{ghi}$

β_0: constant offset or the mean of group 1 if β_1=0

β_g: mean difference to the offset (or group 1)

α_h: mean difference for the second factor

$(\alpha\beta)_{gh}$: interaction effects between the two factors

ϵ_{ghi}: additive error term

Linear Models

following hypotheses are often tested:

i. additive model with the hypothesis H_0: $(\alpha\beta)_{gh} = 0$ for all g and h:

$$y_{ghi} = \beta_0 + \beta_g + \alpha_h + \epsilon_{ghi}$$

this model should be compared to the full model

ii. factor corresponding to α has no effect:

$$y_{ghi} = \beta_0 + \beta_g + \epsilon_{ghi}$$

this model should be compared to the additive model in (i)

iii. factor corresponding to β has no effect:

$$y_{ghi} = \beta_0 + \alpha_h + \epsilon_{ghi}$$

this model should be compared to the additive model in (i)

Linear Models

tested with sum-zero constraints:

i. $\sum_{g=1}^{G} \beta_g = 0$

ii. $\sum_{h=1}^{H} \alpha_h = 0$

iii. $\forall_g : \quad \sum_{h=1}^{H} (\alpha\beta)_{gh} = 0$

iv. $\forall_h : \quad \sum_{g=1}^{G} (\alpha\beta)_{gh} = 0$

OR with corner point constraints:

i. $\beta_1 = 0$

ii. $\alpha_1 = 0$

iii. $\forall_g : \quad (\alpha\beta)_{g1} = 0$

iv. $\forall_h : \quad (\alpha\beta)_{1h} = 0$

Linear Models

β_0: 1
β_g: G
α_h: H
$(\alpha\beta)_{gh}$: GH

$GH+G+H+1=(G+1)(H+1)$ parameters

minimal data set (noise free): GH observations (each factor combination)

both sets of constraints have $G+H+2$ equations:

- corner point constraints use $(\alpha\beta)_{11}$ twice
- sum-zero constraints →last equation follows from other equations

→ both sets of constraints use up $G+H+1$ degrees of freedom

From $\forall_g : \sum_{h=1}^{H}(\alpha\beta)_{gh} = 0$ follows that $\sum_{g=1}^{G}\sum_{h=1}^{H}(\alpha\beta)_{gh} = 0$. We have $\sum_{h=1}^{H}(\sum_{g=1}^{G}(\alpha\beta)_{gh}) = 0$ and $\sum_{g=1}^{G}(\alpha\beta)_{gh} = 0$ for $h < H$ since the last equation is not used. Thus, $\sum_{g=1}^{G}(\alpha\beta)_{gH} = 0$, which is the last equation.

Linear Models

design matrix X should have at least rank GH to distinguish all interaction effects $(\alpha\beta)_{gh}$

simplify notations:

i. mean of group combination: $\bar{y}_{gh} = \frac{1}{n_{gh}} \sum_{i=1}^{n_{gh}} y_{ghi}$

ii. mean of group g: $\bar{y}_{g.} = \frac{1}{\sum_{h=1}^{H} n_{gh}} \sum_{h=1}^{H} \sum_{i=1}^{n_{gh}} y_{ghi}$

iii. mean of group h: $\bar{y}_{.h} = \frac{1}{\sum_{g=1}^{G} n_{gh}} \sum_{g=1}^{G} \sum_{i=1}^{n_{gh}} y_{ghi}$

iv. overall mean: $\bar{y}_{..} = \frac{1}{\sum_{g,h=1,1}^{G,H} n_{gh}} \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{i=1}^{n_{gh}} y_{ghi}$

Linear Models

normal equations: $\boldsymbol{X}^T \boldsymbol{X} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \\ (\boldsymbol{\alpha\beta}) \end{pmatrix} = \boldsymbol{X}^T \boldsymbol{y}$ β_0 first component of β

$\boldsymbol{X}^T \boldsymbol{X}$ not invertible but normal equations for solution $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\alpha}}^T, (\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\beta}})^T)^T$:

$$\left(\sum_{g,h=1,1}^{G,H} n_{gh}\right) \hat{\beta}_0 + \sum_{g=1}^{G} \left(\sum_{h=1}^{H} n_{gh}\right) \hat{\beta}_g + \sum_{h=1}^{H} \left(\sum_{g=1}^{G} n_{gh}\right) \hat{\alpha}_h + \sum_{g=1}^{G}\sum_{h=1}^{H} n_{gh}(\hat{\alpha}\hat{\beta})_{gh} = \sum_{g,h=1,1}^{G,H} n_{gh}\, \bar{y}_{..}$$

$$\left(\sum_{h=1}^{H} n_{gh}\right) \hat{\beta}_0 + \left(\sum_{h=1}^{H} n_{gh}\right) \hat{\beta}_g + \sum_{h=1}^{H} n_{gh}\, \hat{\alpha}_h + \sum_{h=1}^{H} n_{gh}(\hat{\alpha}\hat{\beta})_{gh} = \sum_{h=1}^{H} n_{gh}\, \bar{y}_{g.},\ 1 \le g \le G$$

$$\left(\sum_{g=1}^{G} n_{gh}\right) \hat{\beta}_0 + \sum_{g=1}^{G} n_{gh}\, \hat{\beta}_g + \left(\sum_{g=1}^{G} n_{gh}\right) \hat{\alpha}_h + \sum_{g=1}^{G} n_{gh}(\hat{\alpha}\hat{\beta})_{gh} = \sum_{g=1}^{G} n_{gh}\, \bar{y}_{.h},\ 1 \le h \le H$$

$$n_{gh}\, \hat{\beta}_0 + n_{gh}\, \hat{\beta}_g + n_{gh}\, \hat{\alpha}_h + n_{gh}(\hat{\alpha}\hat{\beta})_{gh} = n_{gh}\bar{y}_{gh},\ 1 \le g \le G,\ 1 \le h \le H$$

plus the zero sum conditions:

$$\sum_{g=1}^{G} \hat{\beta}_g = 0\ ,\ \sum_{h=1}^{H} \hat{\alpha}_h = 0$$

$$\sum_{g=1}^{G} (\hat{\alpha}\hat{\beta})_{gh} = 0\ ,\ \sum_{h=1}^{H} (\hat{\alpha}\hat{\beta})_{gh} = 0$$

GH observations and GH free parameters
→ normal equations can be solved

Linear Models

balanced case: same number of replicates combination of conditions

$$n_{gh} = \tilde{n}$$

means simplify to:

i. mean of group combination gh: $\bar{y}_{gh} = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} y_{ghi}$

ii. mean of group g: $\bar{y}_{g.} = \frac{1}{H\ \tilde{n}} \sum_{h=1}^{H} \sum_{i=1}^{\tilde{n}} y_{ghi}$

iii. mean of group h: $\bar{y}_{.h} = \frac{1}{G\ \tilde{n}} \sum_{g=1}^{G} \sum_{i=1}^{\tilde{n}} y_{ghi}$

iv. overall mean: $\bar{y}_{..} = \frac{1}{G\ H\ \tilde{n}} \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{i=1}^{\tilde{n}} y_{ghi}$

Linear Models

normal equations become:

$$G\,H\,\tilde{n}\,\hat{\beta}_0 + H\,\tilde{n}\sum_{g=1}^{G}\hat{\beta}_g + G\,\tilde{n}\sum_{h=1}^{H}\hat{\alpha}_h + \tilde{n}\sum_{g=1}^{G}\sum_{h=1}^{H}(\hat{\alpha}\hat{\beta})_{gh} = G\,H\,\tilde{n}\,\bar{y}_{..}$$

$$H\,\tilde{n}\,\hat{\beta}_0 + H\,\tilde{n}\,\hat{\beta}_g + \tilde{n}\sum_{h=1}^{H}\hat{\alpha}_h + \tilde{n}\sum_{h=1}^{H}(\hat{\alpha}\hat{\beta})_{gh} = H\,\tilde{n}\,\bar{y}_{g.},\quad 1 \le g \le G$$

$$G\,\tilde{n}\,\hat{\beta}_0 + \tilde{n}\sum_{g=1}^{G}\hat{\beta}_g + G\,\tilde{n}\,\hat{\alpha}_h + \tilde{n}\sum_{g=1}^{G}(\hat{\alpha}\hat{\beta})_{gh} = G\,\tilde{n}\,\bar{y}_{.h},\quad 1 \le h \le H$$

$$\tilde{n}\,\hat{\beta}_0 + \tilde{n}\,\hat{\beta}_g + \tilde{n}\,\hat{\alpha}_h + \tilde{n}\,(\hat{\alpha}\hat{\beta})_{gh} = \tilde{n}\,\bar{y}_{gh},\quad 1 \le g \le G,\quad 1 \le h \le H$$

zero sum conditions

$$\sum_{g=1}^{G}\hat{\beta}_g = 0\ ,\quad \sum_{h=1}^{H}\hat{\alpha}_h = 0$$

$$\sum_{g=1}^{G}(\hat{\alpha}\hat{\beta})_{gh} = 0\ ,\quad \sum_{h=1}^{H}(\hat{\alpha}\hat{\beta})_{gh} = 0$$

Linear Models

normal equations further simplify to

$$G\ H\ \tilde{n}\ \hat{\beta}_0 = G\ H\ \tilde{n}\ \bar{y}_{..}$$

$$H\ \tilde{n}\ \hat{\beta}_0 + H\ \tilde{n}\ \hat{\beta}_g = H\ \tilde{n}\ \bar{y}_{g.},\quad 1 \leq g \leq G$$

$$G\ \tilde{n}\ \hat{\beta}_0 + G\ \tilde{n}\ \hat{\alpha}_h = G\ \tilde{n}\ \bar{y}_{.h},\quad 1 \leq h \leq H$$

$$\tilde{n}\ \hat{\beta}_0 + \tilde{n}\ \hat{\beta}_g + \tilde{n}\ \hat{\alpha}_h + \tilde{n}\ (\hat{\alpha}\hat{\beta})_{gh} = \tilde{n}\ \bar{y}_{gh},\quad 1 \leq g \leq G,\quad 1 \leq h \leq H$$

which gives

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y}_{..} \\
\hat{\beta}_g &= \bar{y}_{g.} - \hat{\beta}_0 = \bar{y}_{g.} - \bar{y}_{..},\quad 1 \leq g \leq G \\
\hat{\alpha}_h &= \bar{y}_{.h} - \hat{\beta}_0 = \bar{y}_{.h} - \bar{y}_{..},\quad 1 \leq h \leq H \\
(\hat{\alpha}\hat{\beta})_{gh} &= \bar{y}_{gh} - \hat{\beta}_0 - \hat{\beta}_g - \hat{\alpha}_h \\
&= \bar{y}_{gh} - \bar{y}_{g.} - \bar{y}_{.h} + \bar{y}_{..}
\end{aligned}$$

These are unbiased estimators for the means → intuitively

Examples

Dried Plant Weights

Genetically similar seeds are randomly assigned to be raised in:

1. nutritionally enriched environment (treatment group) or
2. standard conditions (control group)

using a completely randomized experimental design.

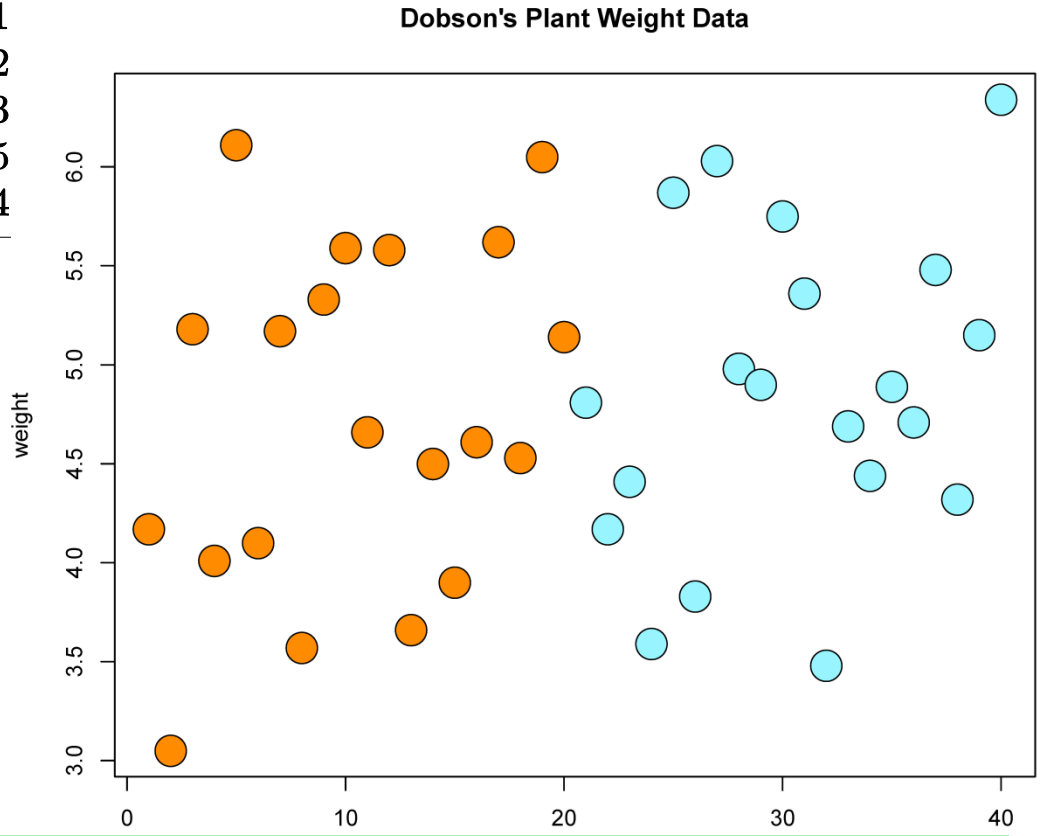
After a predetermined time, all plants are harvested, dried and weighed.

Results (**response**): weight in grams for 20 plants in each group

Goal: test if there is a difference in yield between treatment and control

Linear Models

Treatment group		Control group	
4.81	5.36	4.17	4.66
4.17	3.48	3.05	5.58
4.41	4.69	5.18	3.66
3.59	4.44	4.01	4.50
5.87	4.89	6.11	3.90
3.83	4.71	4.10	4.61
6.03	5.48	5.17	5.62
4.98	4.32	3.57	4.53
4.90	5.15	5.33	6.05
5.75	6.34	5.59	5.14



Linear Models

simple summary of the data:

```
summary(ctl)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.050   4.077   4.635   4.726   5.392   6.110
summary(trt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.480   4.388   4.850   4.860   5.390   6.340
```

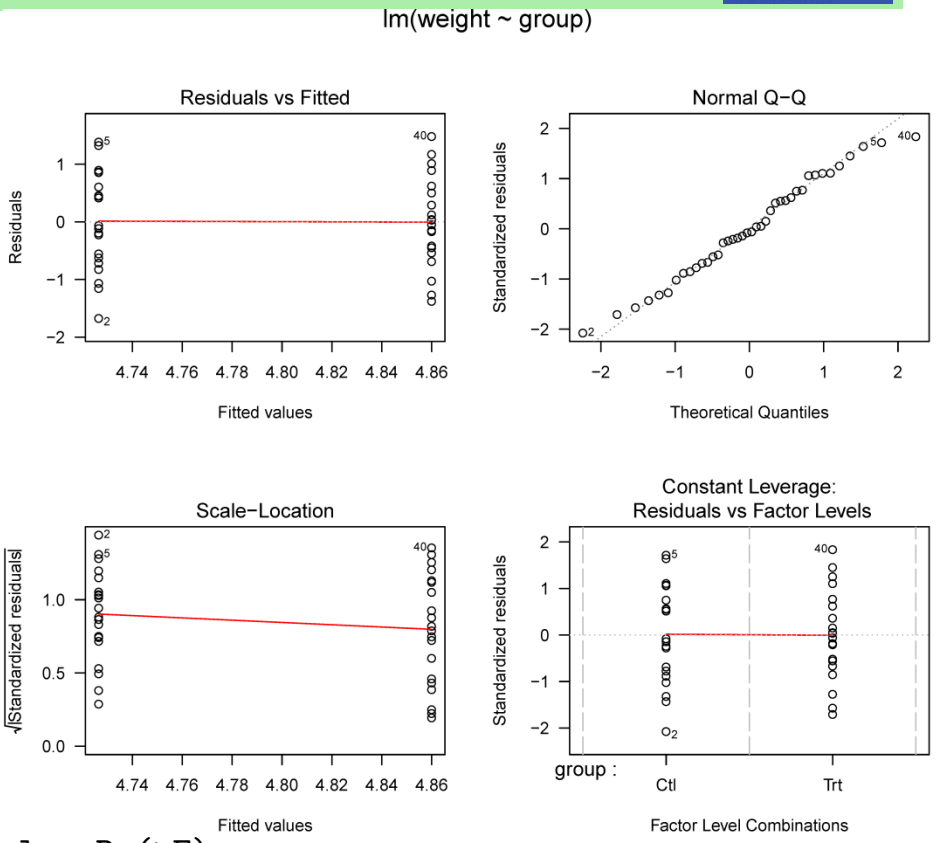
treatment has larger median and larger mean

Is this significant?

When looking at the data in there could be some doubts

→ fit a linear model and print the ANOVA table

Linear Models



fit a linear model:

```
lm.D9 <- lm(weight ~ group)

anova(lm.D9)
Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value Pr(>F)
group      1  0.1782 0.17822  0.2599 0.6131
Residuals 38 26.0535 0.68562
```

linear model:

```
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(lm.D9, las = 1)      # Residuals, Fitted, ...
par(opar)
```

Linear Models

fit a model without an intercept:

```
lm.D90 <- lm(weight ~ group - 1) # omitting intercept
summary(lm.D90)

Call:
lm(formula = weight ~ group - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.67650 -0.57400 -0.05825  0.60763  1.48000

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
groupCtl   4.7265     0.1852   25.53   <2e-16 ***
groupTrt   4.8600     0.1852   26.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.828 on 38 degrees of freedom
Multiple R-squared:  0.9724,    Adjusted R-squared:  0.971
F-statistic: 670.3 on 2 and 38 DF,  p-value: < 2.2e-16
```

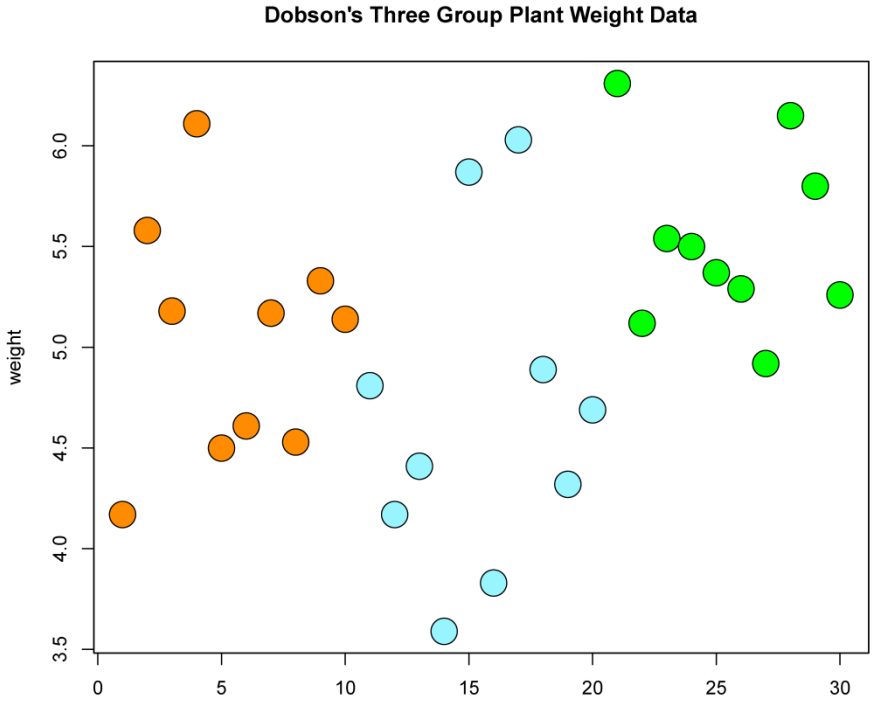
The intercept is replaced by the both groups are significantly different from zero, however there is no difference between the groups.

Linear Models

Extended Dried Plants

The second example extends the first example by another group:

1. control
2. treatment A
3. treatment B



	Control	Treatment A	Treatment B
	4.17	4.81	6.31
	5.58	4.17	5.12
	5.18	4.41	5.54
	6.11	3.59	5.50
	4.50	5.87	5.37
	4.61	3.83	5.29
	5.17	6.03	4.92
	4.53	4.89	6.15
	5.33	4.32	5.80
	5.14	4.69	5.26
$\sum_i y_i$	50.32	46.61	55.26
$\sum_i y_i^2$	256.27	222.92	307.13

Plants from treatment B group (green) seem to be larger

→ check by fitting a linear model

Linear Models


```
anova(lmwg <- lm(weight~group))
Analysis of Variance Table
Response: weight
          Df  Sum Sq Mean Sq F value  Pr(>F)
group      2  3.7663  1.8832  4.8461 0.01591 *
Residuals 27 10.4921  0.3886

summary(lmwg)
Call:
lm(formula = weight ~ group)
Residuals:
    Min      1Q  Median      3Q     Max
-1.0710 -0.4180 -0.0060  0.2627  1.3690

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0320     0.1971  25.527   <2e-16 ***
groupA       -0.3710     0.2788  -1.331   0.1944
groupB        0.4940     0.2788   1.772   0.0877 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear Models

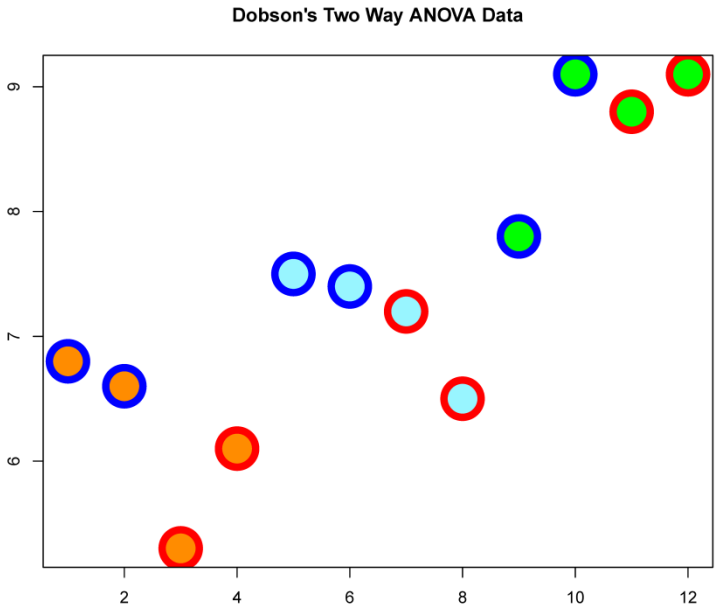

```
Residual standard error: 0.6234 on 27 degrees of freedom
Multiple R-squared:  0.2641,    Adjusted R-squared:  0.2096
F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591

coef(lmwg)
(Intercept)      groupA      groupB
      5.032      -0.371       0.494

coef(summary(lmwg))#- incl.  std.err,  t- and P- values.
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept)    5.032  0.1971284 25.526514 1.936575e-20
groupA        -0.371  0.2787816 -1.330791 1.943879e-01
groupB         0.494  0.2787816  1.771996 8.768168e-02
```

- Group B can be distinguished best from other groups.
  Its coefficient has a p-value of 0.09 which is almost significant.
- The F-statistic and its p-value of 0.016 shows that the groups together are significant.
- The estimated parameters show that group B is larger (0.494) and group A smaller (-0.371) than the control group.

Linear Models

Two-Factor ANOVA Toy Example

In this toy data factor A has 3 levels and factor B has 2 levels.
This gives 6 subgroups with all combinations of A and B levels.
Each subgroup has 2 replicates.



	Levels of factor B		
Levels of factor A	B_1	B_2	Total
A_1	6.8, 6.6	5.3, 6.1	24.8
A_2	7.5, 7.4	7.2, 6.5	28.6
A_3	7.8, 9.1	8.8, 9.1	34.8
Total	45.2	43.0	88.2

Levels of factor A are indicated by the interior color of the circles while levels of factor B are indicated by the border color of the circles.

are there interaction effects?
are there different responses for different levels of factor A?
are there different responses for different levels of factor B?

Linear Models


```
anova(z <- lm(y~a*b))

Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value   Pr(>F)
a          2 12.7400  6.3700 25.8243 0.001127 **
b          1  0.4033  0.4033  1.6351 0.248225
a:b        2  1.2067  0.6033  2.4459 0.167164
Residuals  6  1.4800  0.2467
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The is no evidence against the hypothesis that the levels of factor B do not influence the response.
- Similar there is no evidence against the hypothesis that the interaction effect does not influence the response.
- Therefore we conclude that the response is mainly affected by differences in the levels of factor A.

Analysis of Covariance
The Model

models use covariates: variables measured together with y
analysis of covariance (ANCOVA) models

covariates reduce variance before comparing the means of groups

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

$X\beta$: same as in the ANOVA model (X contains zeros and ones)
Z: covariate values (contains measured values)
u: covariate coefficients

one-way balanced model with one covariate

$$y_{gi} = \beta_0 + \beta_g + u\, z_{gi} + \epsilon_{gi}\,, 1 \leq g \leq G,\ 1 \leq i \leq \tilde{n}$$

β_g: treatment effect
z_{gi}: covariate
u: slope / coefficient for z_{gi}

Linear Models

q covariates: $y_{gi} = \beta_0 + \beta_g + \sum_r^q u_r \, z_{gir} + \epsilon_{gi}, 1 \leq g \leq G, \; 1 \leq i \leq \tilde{n}$

$$\boldsymbol{Zu} = \begin{pmatrix} z_{111} & z_{112} & \cdots & z_{11q} \\ z_{121} & z_{122} & \cdots & z_{12q} \\ \vdots & \vdots & & \vdots \\ z_{G\tilde{n}1} & z_{G\tilde{n}2} & \cdots & z_{G\tilde{n}q} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_q \end{pmatrix}$$

matrices X and Z combined:

$$\boldsymbol{y} = (\boldsymbol{X}, \boldsymbol{Z}) \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{pmatrix} + \boldsymbol{\epsilon}$$

normal equations:

$$\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{Z}^T \end{pmatrix} (\boldsymbol{X}, \boldsymbol{Z}) \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{y} \end{pmatrix}$$

Linear Models

Gives two equations:

$$\boldsymbol{X}^T\boldsymbol{X}\,\hat{\boldsymbol{\beta}} + \boldsymbol{X}^T\boldsymbol{Z}\,\hat{\boldsymbol{u}} = \boldsymbol{X}^T\boldsymbol{y}$$

$$\boldsymbol{Z}^T\boldsymbol{X}\,\hat{\boldsymbol{\beta}} + \boldsymbol{Z}^T\boldsymbol{Z}\,\hat{\boldsymbol{u}} = \boldsymbol{Z}^T\boldsymbol{y}$$

Solving the first equation:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{y} - (\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{Z}\,\hat{\boldsymbol{u}} = \hat{\boldsymbol{\beta}}_0 - (\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{Z}\,\hat{\boldsymbol{u}}$$

$(\boldsymbol{X}^T\boldsymbol{X})^+$ pseudo inverse of $(\boldsymbol{X}^T\boldsymbol{X})$ $\qquad \hat{\boldsymbol{\beta}}_0 = (\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{y}$

substitute the $\hat{\boldsymbol{\beta}}$ equation into the second:

$$\boldsymbol{Z}^T\boldsymbol{X}\,\left((\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{y} - (\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{Z}\,\hat{\boldsymbol{u}}\right) + \boldsymbol{Z}^T\boldsymbol{Z}\,\hat{\boldsymbol{u}} = \boldsymbol{Z}^T\boldsymbol{y}$$

we define $\boldsymbol{P} = \boldsymbol{X}\,(\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T$

Solutions:

$$\hat{\boldsymbol{u}} = \left(\boldsymbol{Z}^T(\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T(\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{y}$$

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 - (\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{Z}\,\hat{\boldsymbol{u}} \qquad \hat{\boldsymbol{\beta}}_0 = (\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{y}$$

Linear Models

Hypotheses:

- H_0: $\beta_1 = \beta_2 = \ldots = \beta_G$ (equality of treatment effects)
- H_0: $\boldsymbol{u} = \mathbf{0}$ (slope equal to zero)
- H_0: $u_1 = u_2 = \ldots = u_q$ (equal slopes, homogeneity of slopes)

Linear Models

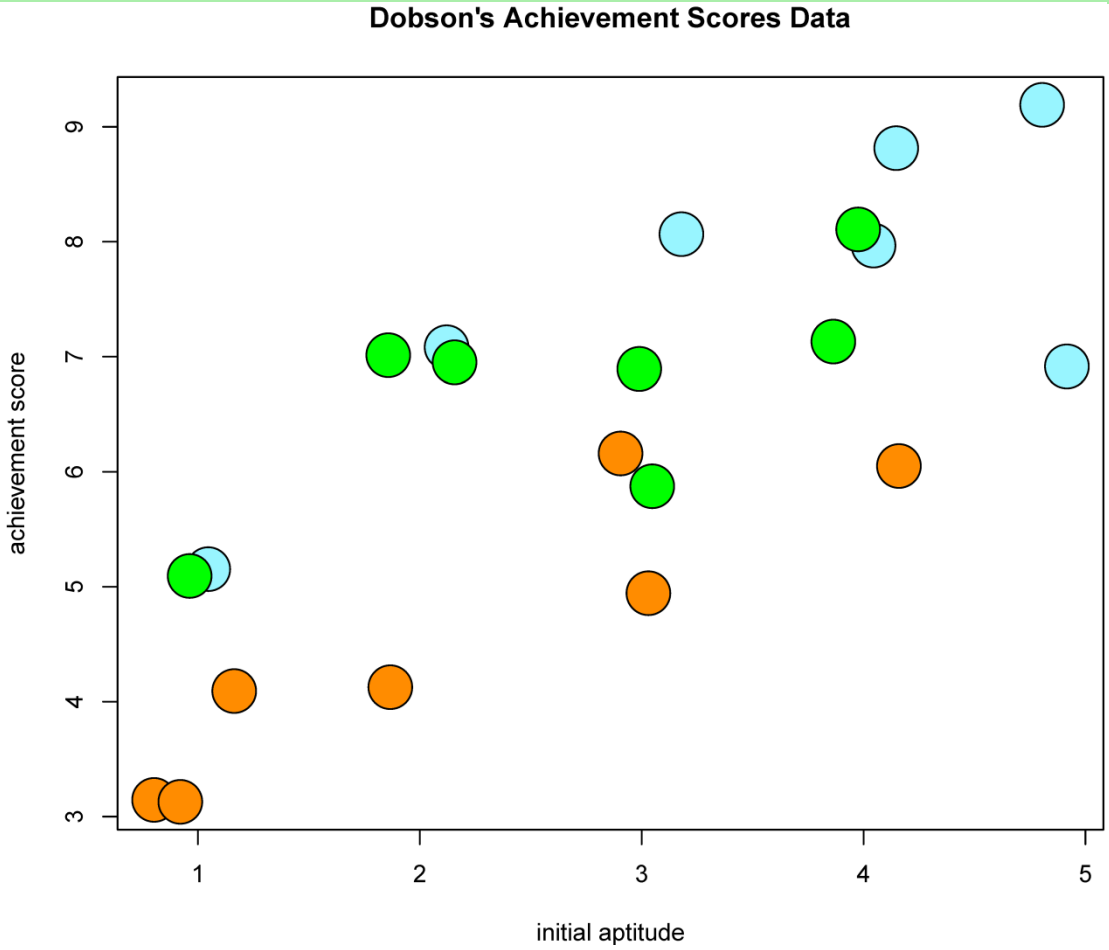
Examples

Achievement Scores

The responses are achievement scores measured at three levels of a factor representing three different training methods. The covariates are aptitude scores measured before training commenced.
Goal: compare the training methods, taking into account differences in initial aptitude between the three groups of subjects.

	Training method					
	A		B		C	
	y	x	y	x	y	x
	6	3	8	4	6	3
	4	1	9	5	7	2
	5	3	7	5	7	2
	3	1	9	4	7	3
	4	2	8	3	8	4
	3	1	5	1	5	1
	6	4	7	2	7	4
$\sum x$ / $\sum y$	31	15	53	24	47	19
$\sum x^2$ / $\sum y^2$	147	41	413	96	321	59
$\sum xy$	75		191		132	

Linear Models



achievement scores y increase linearly with aptitude x.

achievement scores y are generally higher for training methods B and C if compared to A.

Hypothesis: no differences in mean achievement scores among the three training methods, after adjustment for initial aptitude.

Linear Models


```
anova(z <- lm(y~x+m))

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 36.575  36.575  60.355 5.428e-07 ***
m          2 16.932   8.466  13.970 0.0002579 ***
Residuals 17 10.302   0.606
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Of course, the initial aptitude x is significant for the achievement scores y.

More importantly, the training methods, which are given by m, show significant differences concerning the achievement scores.

Linear Models

ANOVA table of different models:

```
z0 <- lm(y~x)
anova(z,z0)
Analysis of Variance Table

Model 1: y ~ x + m
Model 2: y ~ x
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     17 10.302
2     19 27.234 -2   -16.932 13.97 0.0002579 ***
```

Again we see that the training methods show significant differences after adjusting for the initial aptitude.

Linear Models

Birthweights of Girls and Boys

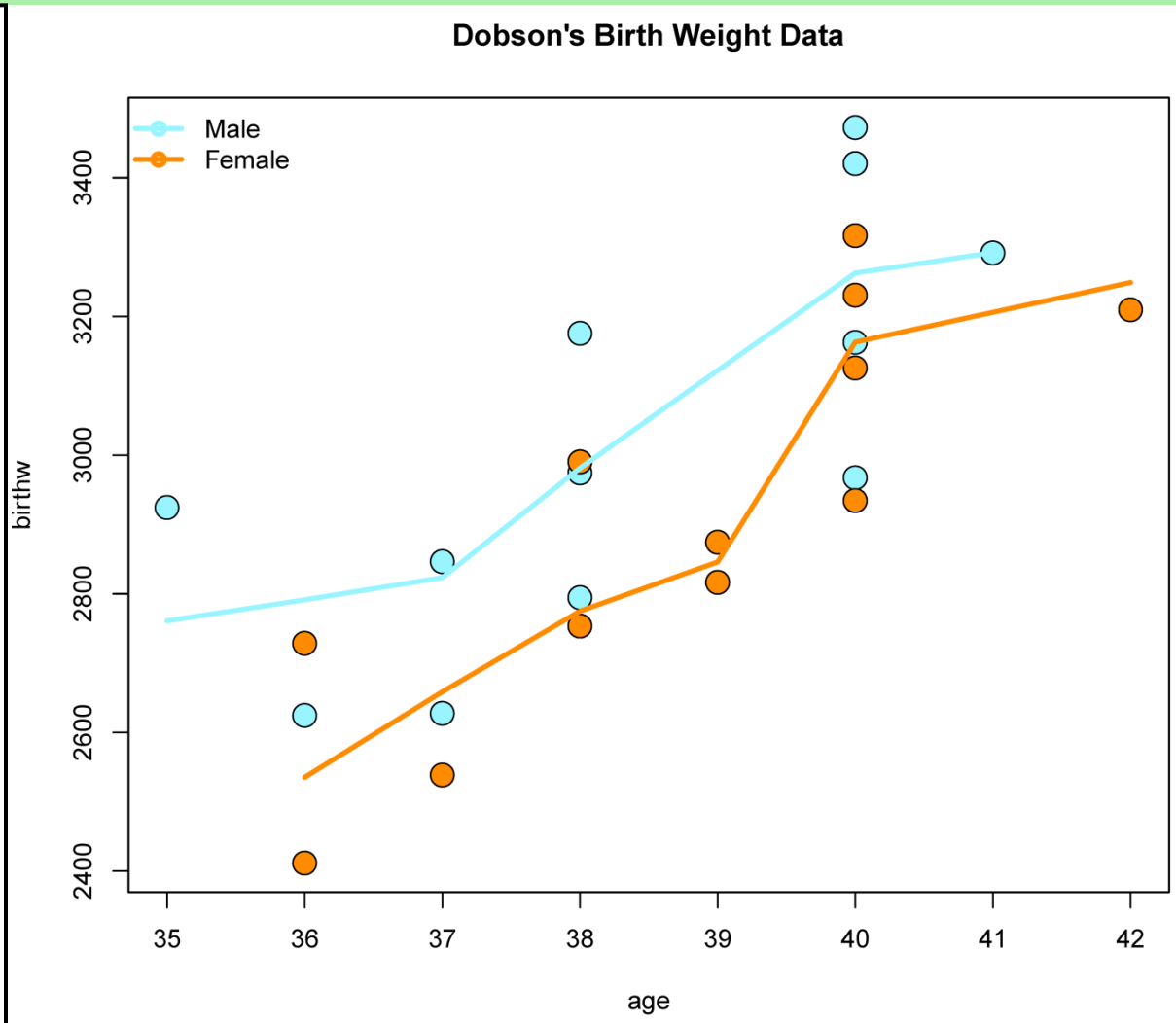
Birthweights in grams and estimated gestational ages in weeks of 12 male and female babies are sampled.

	Boys		Girls	
	Age	Birthweight	Age	Birthweight
	40	2968	40	3317
	38	2795	36	2729
	40	3163	40	2935
	35	2925	38	2754
	36	2625	42	3210
	37	2847	39	2817
	41	3292	40	3126
	40	3473	37	2539
	37	2628	36	2412
	38	3176	38	2991
	40	3421	39	2875
	38	2975	40	3231
Means	38.33	3024.00	38.75	2911.33

mean ages are almost the same for both sexes

mean birthweight for boys is higher than for girls

Linear Models



linear trend of birth weight increasing with gestational age

girls tend to weigh less than boys of the same gestational age

Is the rate of increase of birthweight with gestational age the same for boys and girls?

Linear Models

linear model: groups are male and female; covariate is age

```
summary(l1 <- lm(birthw ~ sex + age), correlation=TRUE)
Call:
lm(formula = birthw ~ sex + age)
Residuals:
    Min      1Q  Median      3Q     Max
-257.49 -125.28  -58.44  169.00  303.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1610.28     786.08  -2.049   0.0532 .
sexFemale    -163.04      72.81  -2.239   0.0361 *
age           120.89      20.46   5.908 7.28e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.‘

Residual standard error: 177.1 on 21 degrees of freedom
Multiple R-squared:   0.64,     Adjusted R-squared:  0.6057
F-statistic: 18.67 on 2 and 21 DF,  p-value: 2.194e-05

Correlation of Coefficients:
          (Intercept) sexFemale
sexFemale  0.07
age       -1.00       -0.12
```

Of course, the birthweight depends on the age, which is highly significant.

However also the sex is significant at a level of 0.05. Females weigh less than males as the coefficient for females is -163.04.

Linear Models

model without an intercept:

```
summary(l0 <- lm(birthw ~ sex + age -1), correlation=TRUE)
Call:
lm(formula = birthw ~ sex + age - 1)
Residuals:
    Min      1Q  Median      3Q     Max
-257.49 -125.28  -58.44  169.00  303.98

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
sexMale   -1610.28     786.08  -2.049   0.0532 .
sexFemale -1773.32     794.59  -2.232   0.0367 *
age         120.89      20.46   5.908 7.28e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 177.1 on 21 degrees of freedom
Multiple R-squared:  0.9969,    Adjusted R-squared:  0.9965
F-statistic:  2258 on 3 and 21 DF,  p-value: < 2.2e-16

Correlation of Coefficients:
          sexMale sexFemale
sexFemale  1.00
age       -1.00   -1.00
```

intercept is now attributed to the males.

previous setting: males were the reference group.

Either the reference group effect or the constant offset (the intercept) is set to zero.

Linear Models

ANOVA table:

```
anova(l1,l0)
Analysis of Variance Table

Model 1: birthw ~ sex + age
Model 2: birthw ~ sex + age - 1
  Res.Df    RSS Df  Sum of Sq F Pr(>F)
1     21 658771
2     21 658771  0 1.5134e-09
```

Linear Models

model which contains the interaction of factor sex with variable age:

```
summary(li <- lm(birthw ~ sex + sex:age -1), correlation=TRUE)
Call:
lm(formula = birthw ~ sex + sex:age - 1)
Residuals:
    Min      1Q  Median      3Q     Max
-246.69 -138.11  -39.13  176.57  274.28
Coefficients
             Estimate Std. Error t value Pr(>|t|)
sexMale       -1268.67    1114.64  -1.138 0.268492
sexFemale     -2141.67    1163.60  -1.841 0.080574 .
sexMale:age     111.98      29.05   3.855 0.000986 ***
sexFemale:age   130.40      30.00   4.347 0.000313 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

Residual standard error: 180.6 on 20 degrees of freedom
Multiple R-squared:  0.9969,    Adjusted R-squared:  0.9963
F-statistic:  1629 on 4 and 20 DF,  p-value: < 2.2e-16

Correlation of Coefficients:
             sexMale sexFemale sexMale:age
sexFemale     0.00
sexMale:age  -1.00    0.00
sexFemale:age  0.00   -1.00      0.00
```

The interaction terms explain significant variance in the data.

Interaction factors are driven by age.

Age less significant: divided into two interaction factors.

Linear Models

ANOVA table:

```
anova(li,l0)
Analysis of Variance Table

Model 1: birthw ~ sex + sex:age - 1
Model 2: birthw ~ sex + age - 1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     20 652425
2     21 658771 -1   -6346.2 0.1945 0.6639
```

The difference between the models is not significant.

Only age is separated into the combined factors containing the sex.

Mixed Effects Models

We now assume there is a second source of variation which is represented by a hidden or latent variable u.
Problem: If the variance of u is not known, then error variance has to be distinguished from the variance through u.

For mixed effect models the variance no longer factors out.
For each observation y there is a latent variable u:

$$y = \boldsymbol{x}^T \boldsymbol{\beta} + \boldsymbol{z}^T \boldsymbol{u} + \epsilon$$

z is a vector indicating the presence of the latent variable, which can be sampled with y or be designed via dummy variables.

Linear Models

Assumptions:

$$\mathrm{E}(\boldsymbol{u}) = \mathbf{0} \;,\; \mathrm{Var}(\boldsymbol{u}) = \boldsymbol{G}$$
$$\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0} \;,\; \mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{R}$$
$$\mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{u}) = \mathbf{0}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$$

Z allows to specify groups or certain measurement conditions.

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$$
$$\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{Z}^T\boldsymbol{G}\boldsymbol{Z} + \boldsymbol{R}$$

Linear Models

Approximative Estimator

estimator for both β and u
estimator for u is the posterior

Estimator for Beta

Assume normal distributed: $\boldsymbol{G} = \sigma_u^2 \boldsymbol{I}$ and $\boldsymbol{R} = \sigma^2 \boldsymbol{I}$

Approximate: $\boldsymbol{G} = \hat{\sigma}_u^2 \boldsymbol{I}$ and $\boldsymbol{R} = \hat{\sigma}^2 \boldsymbol{I}$

estimates using restricted (or residual) maximum likelihood (REML)

$$\boldsymbol{K} = \boldsymbol{C}(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{C}\left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T\right)$$

$\boldsymbol{C}$: is a full-rank transformation of the rows of $(\boldsymbol{I} - \boldsymbol{P})$

$$\boldsymbol{K}\,\boldsymbol{X} = \boldsymbol{0} \qquad \boldsymbol{\Sigma} = \sigma_u^2\,\boldsymbol{Z}\boldsymbol{Z}^T + \sigma^2 \boldsymbol{I}_n$$

$$\boldsymbol{K}\,\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{0}\,,\ \boldsymbol{K}\,\boldsymbol{\Sigma}\,\boldsymbol{K}^T\right)$$

equations by setting the derivatives of the likelihood of $\boldsymbol{Ky}$ with respect to σ^2 and to ${\sigma_u}^2$ to zero:

$$\mathrm{Tr}\left(\boldsymbol{K}^T\left(\boldsymbol{K}\,\boldsymbol{\Sigma}\,\boldsymbol{K}^T\right)^{-1}\boldsymbol{K}\right) = \boldsymbol{y}^T\boldsymbol{K}^T\left(\boldsymbol{K}\,\boldsymbol{\Sigma}\,\boldsymbol{K}^T\right)^{-1}\boldsymbol{K}\,\boldsymbol{K}^T\left(\boldsymbol{K}\,\boldsymbol{\Sigma}\,\boldsymbol{K}^T\right)^{-1}\boldsymbol{K}\,\boldsymbol{y}$$

$$\mathrm{Tr}\left(\boldsymbol{K}^T\left(\boldsymbol{K}\,\boldsymbol{\Sigma}\,\boldsymbol{K}^T\right)^{-1}\boldsymbol{K}\boldsymbol{Z}\boldsymbol{Z}^T\right) = \boldsymbol{y}^T\boldsymbol{K}^T\left(\boldsymbol{K}\,\boldsymbol{\Sigma}\,\boldsymbol{K}^T\right)^{-1}\boldsymbol{K}\,\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{K}^T\left(\boldsymbol{K}\,\boldsymbol{\Sigma}\,\boldsymbol{K}^T\right)^{-1}\boldsymbol{K}\,\boldsymbol{y}$$

Solution of these equations are the estimators for σ^2 and ${\sigma_u}^2$

Using these estimators:

$$\hat{\boldsymbol{\Sigma}} = \hat{\sigma}_u^2\,\boldsymbol{Z}\boldsymbol{Z}^T + \hat{\sigma}^2\boldsymbol{I}_n$$

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\right)^{+}\boldsymbol{X}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}$$

This is the estimated generalized least squares (EGLS) estimator.

The EGLS estimator is only asymptotically the minimum variance unbiased estimator (MVUE).

Linear Models

estimate for the covariance: $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}\right)^+ \boldsymbol{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X} \left(\boldsymbol{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}\right)^+$

Full rank $\boldsymbol{X}$: $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}\right)^{-1}$

Large-sample estimator:

Approximative confidence intervals: $\boldsymbol{a}^T \boldsymbol{\beta} \in \boldsymbol{a}^T \hat{\boldsymbol{\beta}} \pm z_{\alpha/2} \sqrt{\boldsymbol{a}^T \left(\boldsymbol{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}\right)^+ \boldsymbol{a}}$

Small-sample estimator: $t = \dfrac{\boldsymbol{a}^T \hat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{a}^T \left(\boldsymbol{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}\right)^+ \boldsymbol{a}}}$

t-distribution with unknown degrees of freedom

→ task is to estimate the degrees of freedom

Estimator for u

connection between two normally distributed variables:

$$\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_{uu}) \;,\; \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{yy})$$

$$\boldsymbol{\Sigma}_{uv} = \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{u}) \quad \text{and} \quad \boldsymbol{\Sigma}_{vu} = \mathrm{Cov}(\boldsymbol{u}, \boldsymbol{y}):$$

$$\boldsymbol{u} \mid \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_u + \boldsymbol{\Sigma}_{vu}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y) \;,\; \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{vu}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{uv}\right)$$

$$\mathrm{Cov}(\boldsymbol{u}, \boldsymbol{y}) = \mathrm{Cov}(\boldsymbol{u} \;,\; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}) = \boldsymbol{G}\,\boldsymbol{Z}^T$$

$$\mathrm{E}(\boldsymbol{u}) = \boldsymbol{0} \;,\; \mathrm{Var}(\boldsymbol{u}) = \boldsymbol{G}$$

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} \;,\; \mathrm{Var}(\boldsymbol{y}) = \boldsymbol{Z}^T\boldsymbol{G}\boldsymbol{Z} + \boldsymbol{R}$$

Therefore we obtain

$$\boldsymbol{u} \mid \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{G}\,\boldsymbol{Z}^T(\boldsymbol{Z}^T\boldsymbol{G}\boldsymbol{Z} + \boldsymbol{R})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \;,\; \boldsymbol{G} - \boldsymbol{G}\,\boldsymbol{Z}^T(\boldsymbol{Z}^T\boldsymbol{G}\boldsymbol{Z} + \boldsymbol{R})^{-1}\boldsymbol{Z}\,\boldsymbol{G}^T\right)$$

use above approximation:

$$\boldsymbol{G} = \hat{\sigma}_u^2\boldsymbol{I}$$

$$\boldsymbol{R} = \hat{\sigma}^2\boldsymbol{I}$$

$$\boldsymbol{Z}^T\boldsymbol{G}\boldsymbol{Z} + \boldsymbol{R} = \hat{\boldsymbol{\Sigma}}$$

Full Estimator

full estimator and not only an approximation.
Henderson's "mixed model equations" (MME):

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix}$$

Solutions are best linear unbiased estimates (BLUE)

Mixed effect models can also be fitted by the EM algorithm:

- E-step: estimate variance components
- M-step: maximizes parameters

For $\boldsymbol{R} = \sigma^2\boldsymbol{I}$ we obtain for the MME:

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z} + \sigma^{-2}\,\boldsymbol{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{y} \end{pmatrix}$$

Generalized Linear Models

Other distributions than normal errors are possible, even discrete or count distributions.
error-free model as expectation of the observation y_i:
$\mathrm{E}(y_i) = \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$
generalize this relation by introducing a link function g: $g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$

Generalized linear models require:

i. random component or an error distribution which specifies the probability distribution of the response y

ii. systematic component which is a linear function of the explanatory variables / regressors

iii. link function which determines the functional relation between the expectation of the random variable and the systematic component

Linear Models

exponential dispersion model with the natural parameter θ_i and dispersion parameter ϕ, the density is

$$f(y_i \mid \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right)$$

where $b(\theta_i) = a(\phi) \ln \int \exp\left(\frac{y_i\theta_i}{a(\phi)} + c(y_i, \phi)\right) dy_i$

b is a normalizing constant to ensures f to be a distribution:

$$\int f(y_i \mid \theta_i, \phi)\, dy_i = \frac{\int \exp\left(\frac{y_i\theta_i}{a(\phi)} + c(y_i, \phi)\right) dy_i}{\int \exp\left(\frac{y_i\theta_i}{a(\phi)} + c(y_i, \phi)\right) dy_i} = 1$$

$$\mathrm{E}(y_i) = \mu_i = b'(\theta_i)$$
$$\mathrm{Var}(y_i) = b''(\theta_i)\, a(\phi)$$

Linear Models

The log-likelihood is $\ln \mathcal{L} = \sum_{i=1}^{n} \ln \mathcal{L}_i = \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right)$

$\mathcal{L}_i = f(y_i \mid \theta_i, \phi)$ is the conditional likelihood of y_i given x_i

derivative of the log-likelihood with respect to β_j is

$$\frac{\partial \ln \mathcal{L}_i}{\partial \beta_j} = \frac{\partial \ln \mathcal{L}_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial g(\mu_i)} \frac{\partial g(\mu_i)}{\partial \beta_j}$$

only apply the chain rule a couple of times
Using $\mu_i = b'(\theta_i)$ we obtain the derivatives:

$$\frac{\partial \ln \mathcal{L}_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i} \right)^{-1} = (b''(\theta_i))^{-1} = \frac{a(\phi)}{\mathrm{Var}(y_i)}$$

$$\frac{\partial \mu_i}{\partial g(\mu_i)} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-1}$$

$$\frac{\partial g(\mu_i)}{\partial \beta_j} = x_{ij}$$

Linear Models

Maximum → derivative of the log-likelihood is set to zero

$$\frac{\partial \ln \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\, x_{ij}}{\mathrm{Var}(y_i)} \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-1} = 0$$

numerical methods are used to solve this non-linear equation

probability function is determined by a and b while the link by g

to solve this equation by iteratively re-weighted least squares:

$$w_i = \frac{\left(\frac{\partial \mu_i}{\partial g(\mu_i)} \right)^2}{\mathrm{Var}(y_i)} \qquad \left(\boldsymbol{X}^T \boldsymbol{W}^{(k)} \boldsymbol{X} \right) \boldsymbol{\beta}^{(k+1)} = \left(\boldsymbol{X}^T \boldsymbol{W}^{(k)} \boldsymbol{X} \right) \boldsymbol{\beta}^{(k)} + \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}^{(k)}}$$

$\left(\boldsymbol{X}^T \boldsymbol{W}^{(k)} \boldsymbol{X} \right)$ approximates Fisher information matrix: $\mathcal{F} \approx \boldsymbol{X}^T \boldsymbol{W}^{(k)} \boldsymbol{X}$

If X has full rank:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \left(\boldsymbol{X}^T \boldsymbol{W}^{(k)} \boldsymbol{X} \right)^{-1} \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}^{(k)}}$$

Linear Models

Popular generalized linear models given by distribution & link function:

distribution	link function	link name	support	application
normal	$\boldsymbol{X}\boldsymbol{\beta} = g(\mu) = \mu$	identity	real, $(-\infty, +\infty)$	linear response
exponential	$\boldsymbol{X}\boldsymbol{\beta} = g(\mu) = -\mu^{-1}$	inverse	real, $(0, +\infty)$	exponential response
Gamma	$\boldsymbol{X}\boldsymbol{\beta} = g(\mu) = -\mu^{-1}$	inverse	real, $(0, +\infty)$	exponential response
inv. Gaussian	$\boldsymbol{X}\boldsymbol{\beta} = g(\mu) = -\mu^{-2}$	inv. squared	real, $(0, +\infty)$	
Poisson	$\boldsymbol{X}\boldsymbol{\beta} = g(\mu) = \ln(\mu)$	log	integer, $[0, +\infty)$	count data
Bernoulli	$\boldsymbol{X}\boldsymbol{\beta} = g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$	logit	integer, $[0, 1]$	two classes, occurrence
binomial	$\boldsymbol{X}\boldsymbol{\beta} = g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$	logit	integer, $[0, n]$	two classes, count
categorical	$\boldsymbol{X}\boldsymbol{\beta} = g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$	logit	integer, $[0, K]$	K classes, occurrence
multinomial	$\boldsymbol{X}\boldsymbol{\beta} = g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$	logit	integer, $[0, n]^K$	K classes, count

support of the distribution, link name, typical application
last three: (multinomial) logistic regression
Common link functions: `logit, probit, cauchit, cloglog,identity, log, sqrt, inverse squared, inverse`

`cloglog` is the "complementary log log function": $g(x) = \log(-\log(x))$

It is similar to the `logit` models around 0.5 but differs near 0 or 1.

Linear Models

R functions `glm()` and `glm.fit()` for fitting generalized linear models.

For `glm()` the following models are predefined:

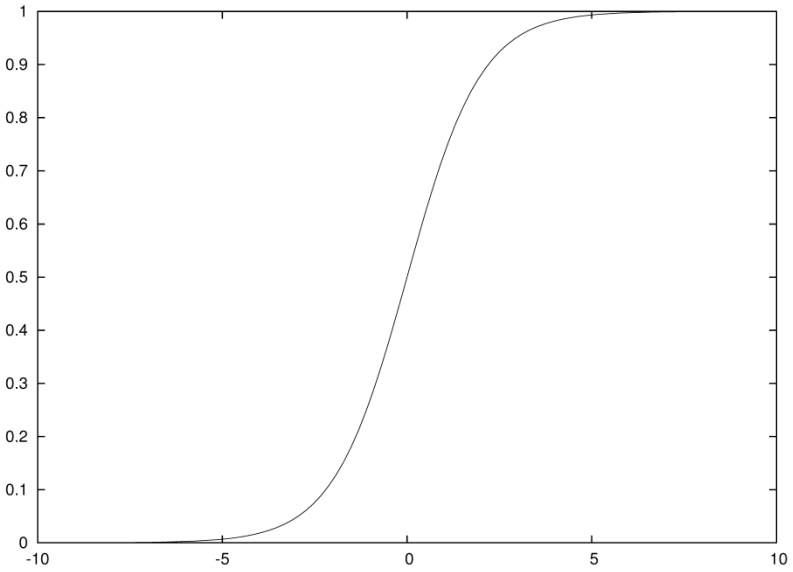
```
binomial(link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
quasi(link = "identity", variance = "constant")
quasibinomial(link = "logit")
quasipoisson(link = "log")
```

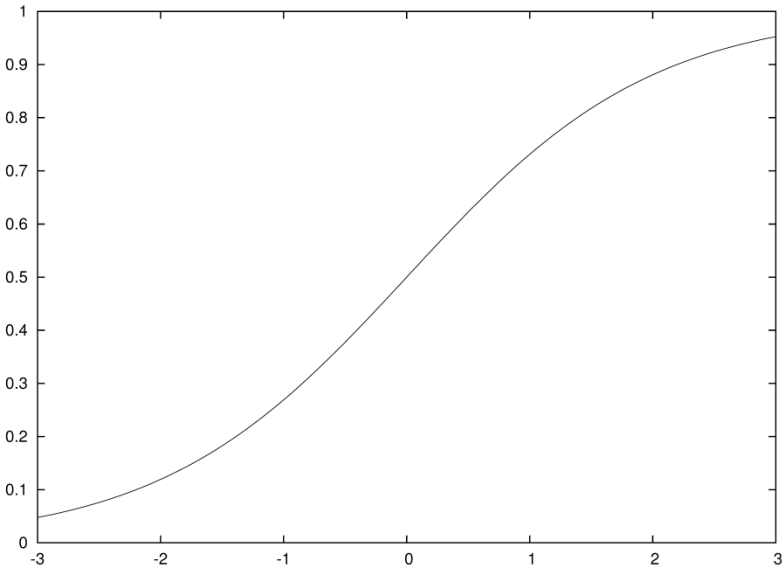
Linear Models

Logistic Regression

inverse of the logit function $g(x) = \ln\left(\frac{x}{1-x}\right)$ is the sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}}$$





Linear Models

$$1 - \frac{1}{1 + e^{-x}} = \frac{e^{-x}}{1 + e^{-x}}$$

gives probabilities

$$p(y = 1 \mid \boldsymbol{x}; \boldsymbol{\beta}) = \frac{1}{1 + e^{-\boldsymbol{x}^T \boldsymbol{\beta}}}$$

$$p(y = 0 \mid \boldsymbol{x}; \boldsymbol{\beta}) = \frac{e^{-\boldsymbol{x}^T \boldsymbol{\beta}}}{1 + e^{-\boldsymbol{x}^T \boldsymbol{\beta}}}$$

logit as link function gives

$$\boldsymbol{x}^T \boldsymbol{\beta} = \ln \left(\frac{p(y = 1 \mid \boldsymbol{x})}{1 - p(y = 1 \mid \boldsymbol{x})} \right)$$

Maximizing the Likelihood

log-likelihood for iid data:

$$\ln \mathcal{L}(\{(y_i, \boldsymbol{x}_i)\}; \boldsymbol{\beta}) = \sum_{i=1}^{n} \ln p(y_i, \boldsymbol{x}_i; \boldsymbol{\beta}) =$$

$$\sum_{i=1}^{n} \ln p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) + \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i)$$

first sum depends on parameter → maximize conditional likelihood:

$$\sum_{i=1}^{n} \ln p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta})$$

Linear Models

derivative of the log-likelihood with $p(y = 1 \mid \boldsymbol{x}_i; \boldsymbol{\beta}) = \frac{1}{1 + e^{- \boldsymbol{x}_i^T \boldsymbol{\beta}}}$:

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \ln p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) =$$

$$\sum_{i=1}^{n} y_i \frac{\partial}{\partial \beta_j} \ln p(y = 1 \mid \boldsymbol{x}_i; \boldsymbol{\beta}) + \sum_{i=1}^{n} (1 - y_i) \frac{\partial}{\partial \beta_j} \ln p(y = 0 \mid \boldsymbol{x}_i; \boldsymbol{\beta}) =$$

$$\sum_{i=1}^{n} - y_i \, p(y = 0 \mid \boldsymbol{x}_i; \boldsymbol{\beta}) x_{ij} + \sum_{i=1}^{n} (1 - y_i) \, p(y = 1 \mid \boldsymbol{x}_i; \boldsymbol{\beta}) x_{ij} =$$

$$\sum_{i=1}^{n} (- y_i \, (1 - p(y = 1 \mid \boldsymbol{x}_i; \boldsymbol{\beta})) \, (1 - y_i) \, p(y = 1 \mid \boldsymbol{x}_i; \boldsymbol{\beta})) \, x_{ij} =$$

$$\sum_{i=1}^{n} (p(y = 1 \mid \boldsymbol{x}_i; \boldsymbol{\beta}) - y_i) \, x_{ij}$$

derivatives have to be zero: $\forall_j : \sum_{i=1}^{n} (p(y = 1 \mid \boldsymbol{x}_i; \boldsymbol{\beta}) - y_i) \, x_{ij} = 0$

Solution by gradient ascent

Linear Models

Alternative formulation with $y \in +1, -1$

$$p(y = 1 \mid \boldsymbol{x}; \boldsymbol{\beta}) = \frac{1}{1 + e^{- \boldsymbol{x}^T \boldsymbol{\beta}}} \qquad p(y = -1 \mid \boldsymbol{x}; \boldsymbol{\beta}) = \frac{e^{- \boldsymbol{x}^T \boldsymbol{\beta}}}{1 + e^{- \boldsymbol{x}^T \boldsymbol{\beta}}} = \frac{1}{1 + e^{\boldsymbol{x}^T \boldsymbol{\beta}}}$$

$$\rightarrow \quad - \ln p(y = y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) = \ln \left(1 + e^{- y_i \, \boldsymbol{x}_i^T \boldsymbol{\beta}}\right)$$

likelihood: $$\mathcal{L} = - \sum_{i=1}^{n} \ln p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) = \sum_{i=1}^{n} \ln \left(1 + e^{- y_i \, \boldsymbol{x}_i^T \boldsymbol{\beta}}\right)$$

derivatives:

$$\frac{\partial L}{\partial \beta_j} = - \sum_{i=1}^{n} y_i \frac{\partial \, \boldsymbol{x}_i^T \boldsymbol{\beta}}{\partial \beta_j} \frac{e^{- y_i \, \boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{- y_i \, \boldsymbol{x}_i^T \boldsymbol{\beta}}} =$$

$$- \sum_{i=1}^{n} y_i \, x_{ij} \, (1 - p(y_i \mid \boldsymbol{x}; \boldsymbol{\beta}))$$

matrix notation:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = - \sum_{i=1}^{n} y_i \, (1 - p(y_i \mid \boldsymbol{x}; \boldsymbol{\beta})) \, \boldsymbol{x}_i$$

The log likelihood of logistic regression is strictly convex.
→ efficient gradient-based techniques for maximum likelihood

Multinomial Logistic Regression: Softmax

multi-class problems: logistic regression generalized to Softmax

K classes: $y \in \{1, \ldots, K\}$

probability of $\boldsymbol{x}$ belonging to class k: $p(y = k \mid \boldsymbol{x}; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = \frac{e^{\boldsymbol{x}^T \boldsymbol{\beta}_k}}{\sum_{j=1}^{K} e^{\boldsymbol{x}^T \boldsymbol{\beta}_j}}$

Conditional likelihood:

$$L = -\sum_{i=1}^{n} \ln p(y = y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{K} e^{\boldsymbol{x}^T \boldsymbol{\beta}_j} \right) - \boldsymbol{x}^T \boldsymbol{\beta}_{y_i}$$

$$p(y = k \mid \boldsymbol{x}; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = p(k \mid \boldsymbol{x}; \boldsymbol{W}) \qquad \boldsymbol{W} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$$

Linear Models

Derivatives:

$$\frac{\partial \mathcal{L}}{\partial \beta_{kt}} = \sum_{i=1}^{n} \frac{\partial \boldsymbol{x}_i^T \boldsymbol{\beta}_k}{\partial \beta_{kt}} p(k \mid \boldsymbol{x}_i; \boldsymbol{W}) - \delta_{y_i=k} \sum_{i=1}^{n} \frac{\partial \boldsymbol{x}_i^T \boldsymbol{\beta}_k}{\partial \beta_{kt}}$$

$$= \sum_{i=1}^{n} x_{it} p(k \mid \boldsymbol{x}_i; \boldsymbol{W}) - \delta_{y_i=k} \sum_{i=1}^{n} x_{it}$$

Softmax is strictly convex.

Linear Models

Poisson Regression

model **count data**:

1. binomial (variance smaller than the mean)
2. Poisson (variance equal to the mean)
3. negative binomial (variance larger than the mean)

distribution	parameters	pmf $\Pr(X=k)$	μ	Var	$r=\mu/\text{Var}$	r
binomial	$n\in\mathbb{N}$, p	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$	$1/(1-p)$	>1
Poisson	$0<\lambda$	$\frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ	1	$=1$
negative binomial	$0<r$, p	$\binom{k+r-1}{k}(1-p)^r p^k$	$\frac{pr}{1-p}$	$\frac{pr}{(1-p)^2}$	$(1-p)$	<1

$p\in[0,1]$: probability of a success
pmf: probability mass function
μ: mean
Var: variance
$r=\frac{\mu}{\text{Var}}$: ratio of mean to variance

Linear Models

Description of observations:

1. rate θ
2. number of trials n
3. observation: $\lambda = \theta\, n$ number successes or failures out of n trials

depending on the application:

1. rate θ or
2. number of trials or exposures n

changes

Example: n is the number of kilometers which an individual drives with a car, while θ is the probability of having an accident.

i. individuals drove a different number of kilometers → trials changes

ii. all persons drive on a test track 100 km, however, different persons consumed a different amount of alcohol → probability of having an accident, is different for each individual.

Linear Models

Poisson regression models the case were the rate changes:

$$\mathrm{E}(y_i) = \lambda_i = n_i\, \theta_i = n_i\, e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}$$

$$\log \lambda_i = \log n_i + \boldsymbol{x}_i^T \boldsymbol{\beta}$$

term $\log n_i$ is an additional offset

Standard hypotheses tests are possible.

Linear Models

standard error: $\mathrm{SE}(\hat{\boldsymbol{\beta}}) = \sqrt{\frac{1}{\mathcal{F}}}$ $\mathcal{F}$ is the Fisher information matrix

Confidence intervals: $\frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)} \sim \mathcal{N}(0,1)$

Estimated values: $e_i = n_i \, e^{\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}}$ with sd $\sqrt{e_i}$

Pearson residuals: $r_i = \frac{o_i - e_i}{\sqrt{e_i}}$ o_i: observed counts

standardized residuals: $r_{pi} = \frac{o_i - e_i}{\sqrt{e_i}\sqrt{1 - P_{ii}}}$

P_{ii}: leverage; i-th element of the diagonal of the hat matrix P

Linear Models

error is chi-squared distributed: $\sum_i r_i^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$

Poisson regression is a log-linear model: $\log \mathrm{E}(y_i) = c + \boldsymbol{x}_i^T \boldsymbol{\beta}$

Log-linear models:

$$\log \mathrm{E}(y_{jk}) = \log n + \log \theta_{j.} + \log \theta_{.k}$$

$$\log \mathrm{E}(y_{jk}) = \log n + \log \theta_{jk.}$$

$$\log \mathrm{E}(y_{jk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$$

ANOVA like approaches are possible for generalized linear models.

Examples

Birthweight Data: Normal

Our birthweight model `10` was a linear model estimated by least squares. It is a generalized linear model with Gaussian error:

```
summary(zi <- glm(birthw ~ sex + age, family=gaussian()))
Call:
glm(formula = birthw ~ sex + age, family = gaussian())
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-257.49  -125.28   -58.44   169.00   303.98
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1610.28     786.08  -2.049   0.0532 .
sexFemale    -163.04      72.81  -2.239   0.0361 *
age           120.89      20.46   5.908 7.28e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for gaussian family taken to be 31370.04)
    Null deviance: 1829873  on 23  degrees of freedom
Residual deviance:  658771  on 21  degrees of freedom
AIC: 321.39

Number of Fisher Scoring iterations: 2
```

Linear Models

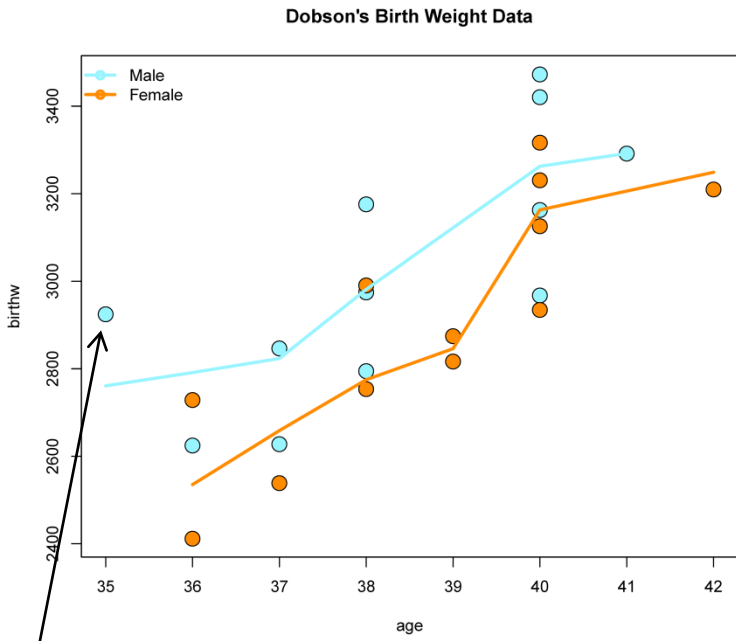
model without intercept:

```
summary(z0 <- glm(birthw ~ sex + age - 1, family=gaussian()))

Call:
glm(formula = birthw ~ sex + age - 1, family = gaussian())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-257.49  -125.28   -58.44   169.00   303.98

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
sexMale   -1610.28     786.08  -2.049   0.0532 .
sexFemale -1773.32     794.59  -2.232   0.0367 *
age         120.89      20.46   5.908 7.28e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 31370.04)

    Null deviance: 213198964  on 24  degrees of freedom
Residual deviance:    658771  on 21  degrees of freedom
AIC: 321.39

Number of Fisher Scoring iterations: 2
```

Linear Models

compare models by an ANOVA table:

```
anova(zi, z0)
Analysis of Deviance Table

Model 1: birthw ~ sex + age
Model 2: birthw ~ sex + age - 1
  Resid. Df Resid. Dev Df    Deviance
1        21     658771
2        21     658771  0 -1.1642e-10
```



observation (35,2925) of a male baby looks like an outlier

check the residuals:

```
z0$residuals
          1           2           3           4           5           6
-257.490545 -188.701891  -62.490545  303.981090 -116.913237  -15.807564
          7           8           9          10          11          12
 -54.384872  247.509455 -234.807564  192.298109  195.509455   -8.701891
         13          14          15          16          17          18
 254.548758  150.126066 -127.451242  -66.662588  -94.239896 -124.556915
         19          20          21          22          23          24
  63.548758 -160.768261 -166.873934  170.337412  -66.556915  168.548758
```

Linear Models

investigate the data by removing the observation no. 4
previous models using update:

```
summary(z.o4 <- update(z0, subset = -4))
Call:
glm(formula = birthw ~ sex + age - 1, family = gaussian(), subset = -4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-253.86  -129.46   -53.46   165.04   251.14

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
sexMale   -2318.03     801.57  -2.892  0.00902 **
sexFemale -2455.44     803.79  -3.055  0.00625 **
age         138.50      20.71   6.688 1.65e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 26925.39)
    Null deviance: 204643339  on 23  degrees of freedom
Residual deviance:    538508  on 20  degrees of freedom
AIC: 304.68

Number of Fisher Scoring iterations: 2
```

Now all regressors are more significant.

Linear Models

interaction term:

```
summary(zz <- update(z0, birthw ~ sex+age-1 + sex:age))
Call:
glm(formula = birthw ~ sex + age + sex:age - 1, family = gaussian())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-246.69  -138.11   -39.13   176.57   274.28

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
sexMale       -1268.67    1114.64  -1.138 0.268492
sexFemale     -2141.67    1163.60  -1.841 0.080574 .
age             111.98      29.05   3.855 0.000986 ***
sexFemale:age    18.42      41.76   0.441 0.663893
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 32621.23)

    Null deviance: 213198964  on 24  degrees of freedom
Residual deviance:    652425  on 20  degrees of freedom
AIC: 323.16

Number of Fisher Scoring iterations: 2
```

Linear Models

interaction does not help

ANOVA table tells the same story:

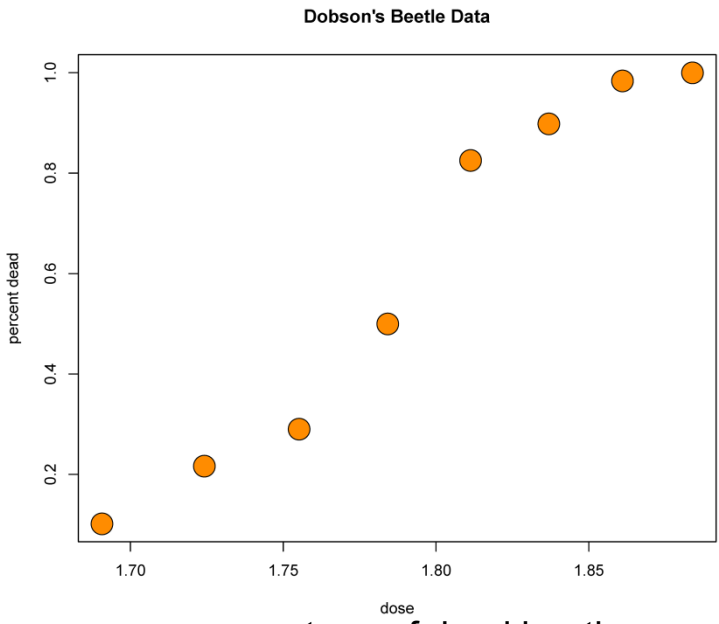
```
anova(z0,zz)
Analysis of Deviance Table

Model 1: birthw ~ sex + age - 1
Model 2: birthw ~ sex + age + sex:age - 1
  Resid. Df Resid. Dev Df Deviance
1        21     658771
2        20     652425  1   6346.2
```

Linear Models

Beetle Mortality: Logistic Regression

The numbers of dead beetles are counted after five hours exposure to gaseous carbon disulfide at various concentrations given as the logarithm of the quantity (Bliss 1935).

Dose ($\log_{10}$ $CS_2 mgl^{-1}$)	Number of beetles	Number killed
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60



percentage of dead beetles

Linear Models

count data as pairs of (dead,alive)
logistic regression: distribution is binomial and the link is logit

```
summary(zlog <- glm(dead ~ dose, family=binomial(link=logit)))
Call:
glm(formula = dead ~ dose, family = binomial(link = logit))
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
dose          34.270      2.912   11.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43
Number of Fisher Scoring iterations: 4
```

Both intercept and dose are significant as mean is not around zero. The significance of the dose shows that the number of dead beetles indeed depends on the dose of carbon disulfide.

Linear Models

link function probit:

```
summary(zprob <- glm(dead ~ dose, family=binomial(link=probit)))
Call:
glm(formula = dead ~ dose, family = binomial(link = probit))
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5714  -0.4703   0.7501   1.0632   1.3449

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -34.935      2.648  -13.19   <2e-16 ***
dose          19.728      1.487   13.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.20  on 7  degrees of freedom
Residual deviance:  10.12  on 6  degrees of freedom
AIC: 40.318
Number of Fisher Scoring iterations: 4
```

The result is very similar to the logit link function.

Linear Models

link function cloglog:

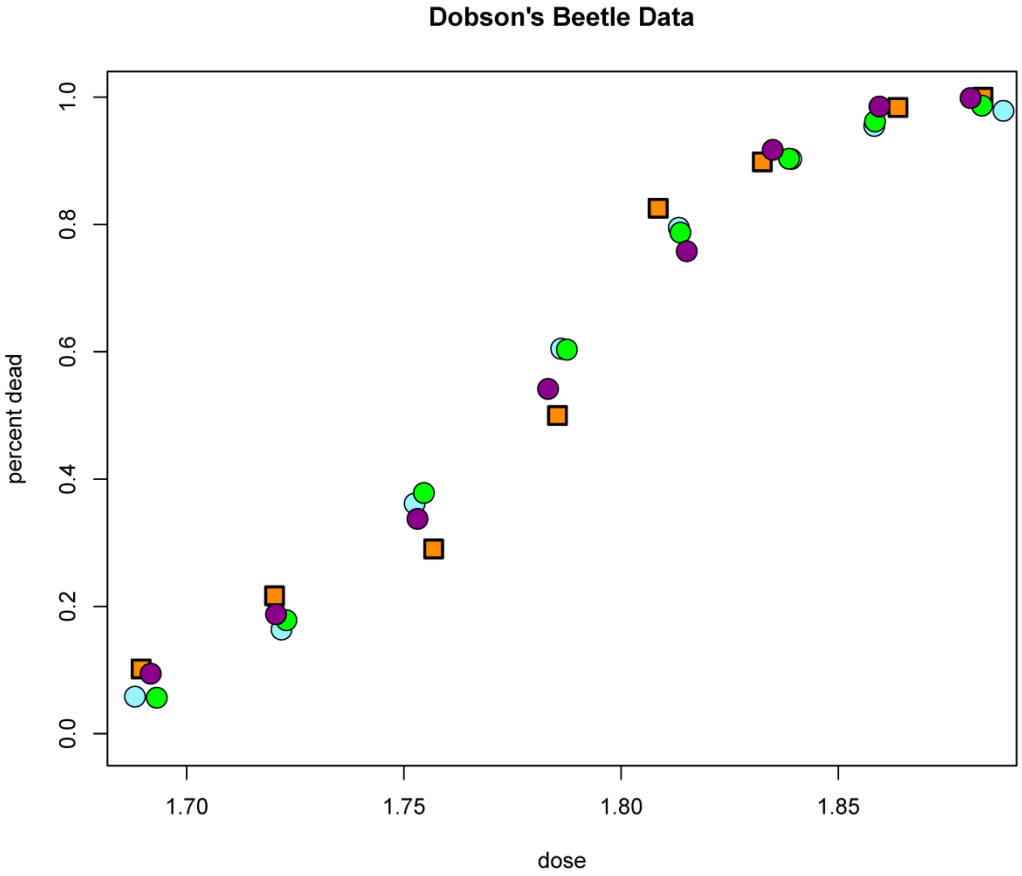
```
summary(zclog <- glm(dead ~ dose, family=binomial(link=cloglog)))
Call:
glm(formula = dead ~ dose, family = binomial(link = cloglog))
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.80329  -0.55135   0.03089   0.38315   1.28883
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -39.572      3.240  -12.21   <2e-16 ***
dose          22.041      1.799   12.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.2024  on 7  degrees of freedom
Residual deviance:   3.4464  on 6  degrees of freedom
AIC: 33.644

Number of Fisher Scoring iterations: 4
```

For cloglog the residual deviance is 3.4464 while it was 11.232 and 10.12 for the logit and probit. AIC (Akaike information criterion) of the last model is lower → last model fits the data better.

Linear Models

fitting of different link functions

Orange rectangles are the original data, blue circles are the fitted points with logistic link function, green circles are the fitted points with the probit link function, and the magenta circles are the fitted points with the cloglog link function.



The x-axis values are jittered.

The cloglog link function fits the points best.

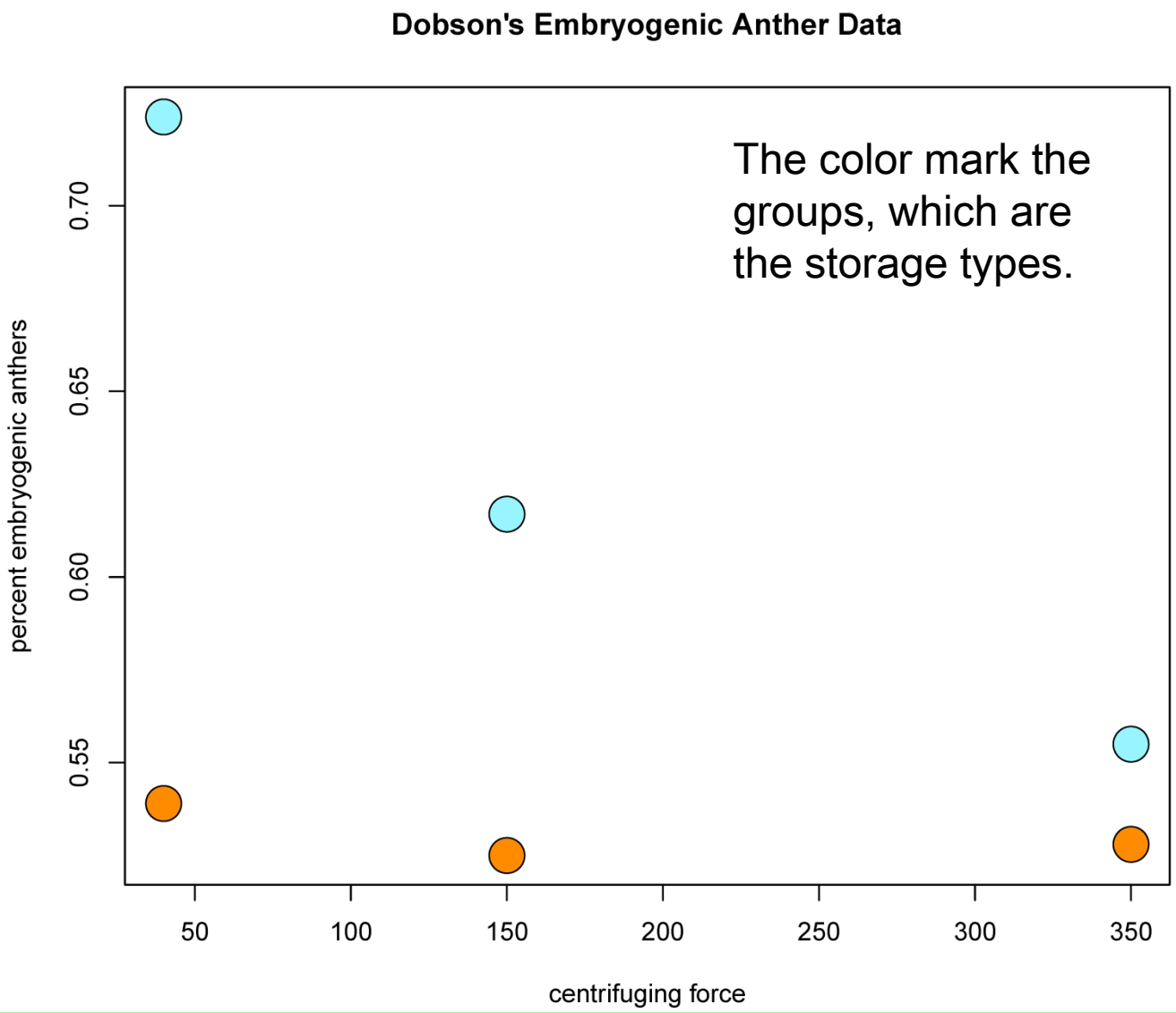
Linear Models

Embryogenic Anthers: Logistic Regression

Data from Sangwan-Norrell (1977) where the authors counted the embryogenic anthers of the plant species *Datura innoxia* Mill. obtained from a particular number of anthers prepared. The embryogenic anthers were obtained under different conditions.
The first factor has two levels which relate to the storage type, which is either a control storage or a storage at 3°C for 48 hours.
The second factor has three levels corresponding to the centrifuging forces.
Goal: compare the treatment and the control storage type after adjusting for the centrifuging force.

		Centrifuging force (g)		
Storage condition		40	150	350
Control	y	55	52	57
	n	102	99	108
Treatment	y	55	50	50
	n	76	81	90

Linear Models



Linear Models

full model:

```
summary(glm(y ~ g*f, family=binomial(link="logit")))
Call:
glm(formula = y ~ g * f, family = binomial(link = "logit"))
Deviance Residuals:
       1         2         3         4         5         6
 0.08269  -0.12998   0.04414   0.42320  -0.60082   0.19522

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.1456719  0.1975451   0.737   0.4609
g2           0.7963143  0.3125046   2.548   0.0108 *
f           -0.0001227  0.0008782  -0.140   0.8889
g2:f        -0.0020493  0.0013483  -1.520   0.1285
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10.45197  on 5  degrees of freedom
Residual deviance:  0.60387  on 2  degrees of freedom
AIC: 38.172

Number of Fisher Scoring iterations: 3
```

Linear Models

No interaction effect between centrifuging force and storage type:

```
summary(glm(y ~ g + f, family=binomial(link="logit")))
Call:
glm(formula = y ~ g + f, family = binomial(link = "logit"))
Deviance Residuals:
      1        2        3        4        5        6
-0.5507  -0.2781   0.7973   1.1558  -0.3688  -0.6584
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.306643   0.167629   1.829   0.0674 .
g2           0.405554   0.174560   2.323   0.0202 *
f           -0.000997   0.000665  -1.499   0.1338
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10.4520  on 5  degrees of freedom
Residual deviance:  2.9218  on 3  degrees of freedom
AIC: 38.49

Number of Fisher Scoring iterations: 3
```

The centrifuging force seems not to be relevant for explaining the yield in embryogenic anthers.

Linear Models

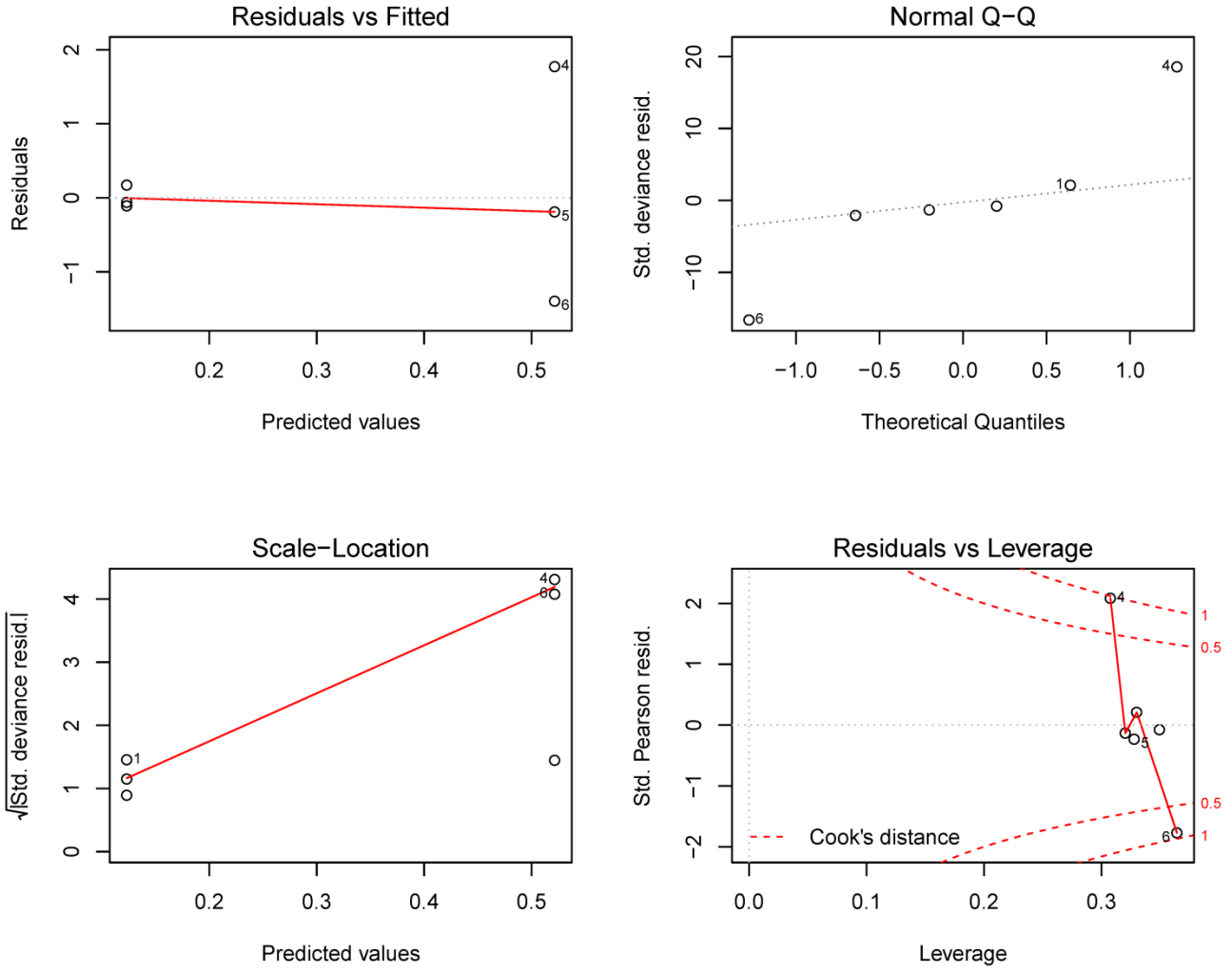
→ we only consider the groups storage conditions:

```
summary(glm.p84 <- glm(y~g,  family=binomial(link="logit")))
Call:
glm(formula = y ~ g, family = binomial(link = "logit"))
Deviance Residuals:
       1         2         3         4         5         6
 0.17150  -0.10947  -0.06177   1.77208  -0.19040  -1.39686
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1231     0.1140   1.080   0.2801
g2            0.3985     0.1741   2.289   0.0221 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 10.452  on 5  degrees of freedom
Residual deviance:  5.173  on 4  degrees of freedom
AIC: 38.741

Number of Fisher Scoring iterations: 3
```

best model with respect to the AIC: which only consider the groups

Linear Models

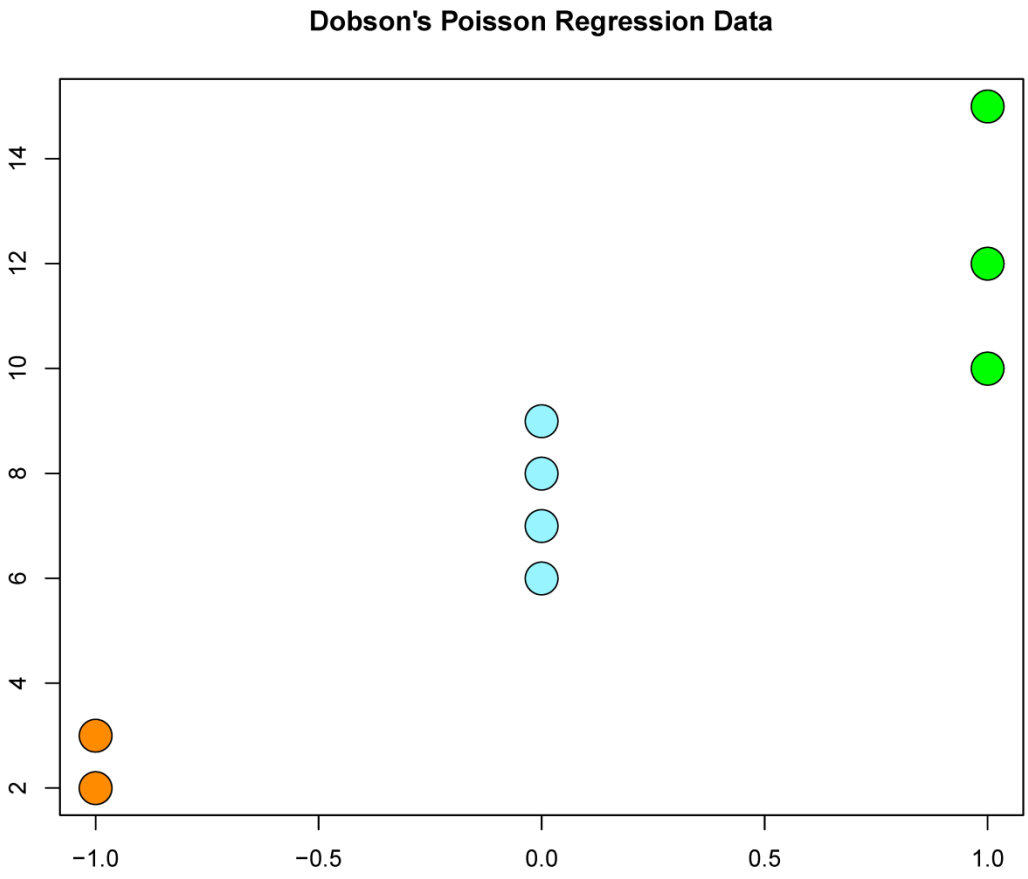
best model which only consider the groups is analyzed



Toy Example 1: Poisson Regression

Poisson regression with toy example from Dobson 2002, page 71.

There is a clear relation between x and the count data y as counts for x=1.0 are larger than counts for x=0.0 which in turn are larger than counts for x=-1.0.



Linear Models


```
summary(glm(y~x, family=poisson(link="identity")))
Call:
glm(formula = y ~ x, family = poisson(link = "identity"))
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7019  -0.3377  -0.1105   0.2958   0.7184
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.4516     0.8841   8.428  < 2e-16 ***
x             4.9353     1.0892   4.531 5.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 18.4206  on 8  degrees of freedom
Residual deviance:  1.8947  on 7  degrees of freedom
AIC: 40.008

Number of Fisher Scoring iterations: 3
```

Both the intercept and the coefficient are significant.
The intercept must move x into the range of the count data.

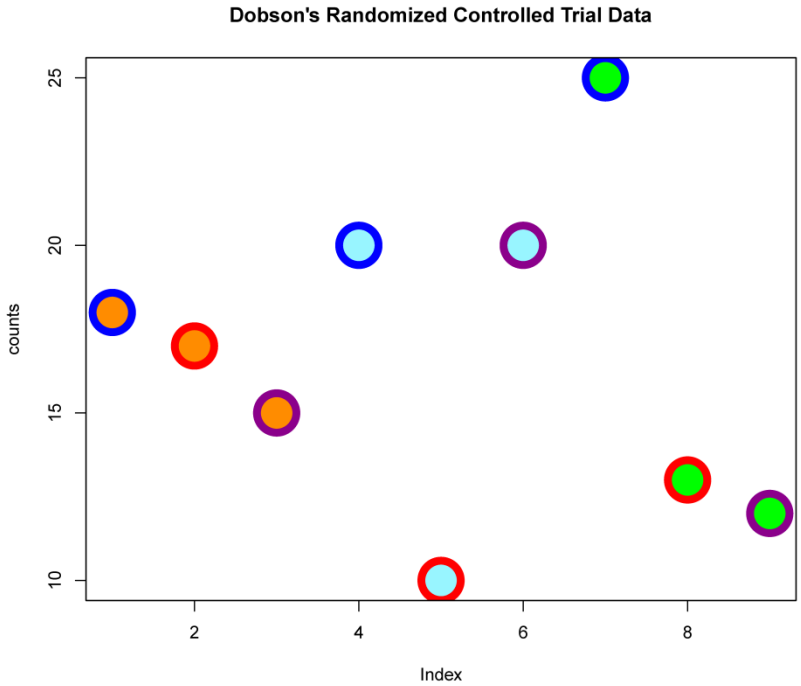
Toy Example 2: Poisson Regression

Another example for Poisson regression from Dobson, 1990, pp. 93. A randomized controlled trial with two factors, outcome and treatment, with three levels. Each treatment group contains 50 samples.

	Outcome			
Treatment	O_1	O_2	O_3	Total
T_1	18	17	15	50
T_2	20	10	20	50
T_3	25	13	12	50
Total	63	40	47	

Outcomes are indicated by the border color of the circles (O_1=blue,O_2=red,O_3=magenta).

Treatments are indicated by the interior color of the circles (T_1=orange,T_2=blue,T_3=green).



The counts for outcome O_1 are larger than the other two.

Linear Models

Poisson regression:

```
summary(z <- glm(counts ~ outcome + treatment, family=poisson()))
Call:
glm(formula = counts ~ outcome + treatment, family = poisson())
Deviance Residuals:
      1         2         3         4         5         6         7
8
-0.67125   0.96272  -0.16965  -0.21999  -0.95552   1.04939   0.84715  -
0.09167
      9
-0.96656
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.045e+00  1.709e-01  17.815   <2e-16 ***
outcome2    -4.543e-01  2.022e-01  -2.247   0.0246 *
outcome3    -2.930e-01  1.927e-01  -1.520   0.1285
treatment2   8.717e-16  2.000e-01   0.000   1.0000
treatment3   4.557e-16  2.000e-01   0.000   1.0000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 10.5814  on 8  degrees of freedom
Residual deviance:  5.1291  on 4  degrees of freedom
AIC: 56.761
Number of Fisher Scoring iterations: 4
```

intercept is significant: data not centered around zero.

Reference: outcome 1 and treatment 1

Treatment: no influence on the counts

Outcome O_2 is significant for a level of 0.01.

Detergent Brand: Poisson Regression

These data were reported by Ries & Smith (1963) and analyzed by Cox & Snell (1989). The user preference for brand M or X is counted. Explanatory variables (regressors, features) are “user of M”, “temperature”, and “water”.

user of M?	No				Yes			
temperature	Low		High		Low		High	
preference	X	M	X	M	X	M	X	M
water softness								
hard	68	42	42	30	37	52	24	43
medium	66	50	33	23	47	55	23	47
soft	63	53	29	27	57	49	19	29

Linear Models

Poisson regression:

```
summary(detg.m0 <- glm(Fr ~ M.user*Temp*Soft + Brand,family=poisson,data = detg))
Call:
glm(formula = Fr ~ M.user * Temp * Soft + Brand, family = poisson, data = detg)
     Min        1Q    Median        3Q       Max
-2.20876  -0.99190  -0.00126   0.93542   1.97601
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.01524    0.10034  40.018  < 2e-16 ***
M.userY      -0.21184    0.14257  -1.486  0.13731
TempHigh     -0.42381    0.15159  -2.796  0.00518 **
SoftMedium    0.05311    0.13308   0.399  0.68984
SoftSoft      0.05311    0.13308   0.399  0.68984
BrandM       -0.01587    0.06300  -0.252  0.80106
M.userY:TempHigh             0.13987    0.22168   0.631  0.52806
M.userY:SoftMedium           0.08323    0.19685   0.423  0.67245
M.userY:SoftSoft             0.12169    0.19591   0.621  0.53449
TempHigh:SoftMedium         -0.30442    0.22239  -1.369  0.17104
TempHigh:SoftSoft           -0.30442    0.22239  -1.369  0.17104
M.userY:TempHigh:SoftMedium  0.21189    0.31577   0.671  0.50220
M.userY:TempHigh:SoftSoft   -0.20387    0.32540  -0.627  0.53098
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 118.627  on 23  degrees of freedom
Residual deviance:  32.826  on 11  degrees of freedom
AIC: 191.24
```

Besides the intercept only temperature is significant but not the water characteristic nor the previous use of the brand.

Linear Models

another model:

```
detg.mod <- glm(terms(Fr ~ M.user*Temp*Soft + Brand*M.user*Temp,
+  keep.order = TRUE), family = poisson, data = detg)

summary(detg.mod, correlation = TRUE, symbolic.cor = TRUE)

Call:
glm(formula = terms(Fr ~ M.user * Temp * Soft + Brand * M.user *
    Temp, keep.order = TRUE), family = poisson, data = detg)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.91365  -0.35585   0.00253   0.33027   0.92146
```

Linear Models


```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.14887    0.10603  39.128  < 2e-16 ***
M.userY      -0.40521    0.16188  -2.503  0.01231 *
TempHigh     -0.44275    0.17121  -2.586  0.00971 **
M.userY:TempHigh          -0.12692    0.26257  -0.483  0.62883
SoftMedium    0.05311    0.13308   0.399  0.68984
SoftSoft      0.05311    0.13308   0.399  0.68984
M.userY:SoftMedium         0.08323    0.19685   0.423  0.67245
M.userY:SoftSoft           0.12169    0.19591   0.621  0.53449
TempHigh:SoftMedium       -0.30442    0.22239  -1.369  0.17104
TempHigh:SoftSoft         -0.30442    0.22239  -1.369  0.17104
M.userY:TempHigh:SoftMedium  0.21189    0.31577   0.671  0.50220
M.userY:TempHigh:SoftSoft   -0.20387    0.32540  -0.627  0.53098
BrandM       -0.30647    0.10942  -2.801  0.00510 **
M.userY:BrandM             0.40757    0.15961   2.554  0.01066 *
TempHigh:BrandM            0.04411    0.18463   0.239  0.81119
M.userY:TempHigh:BrandM    0.44427    0.26673   1.666  0.09579 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 118.627  on 23  degrees of freedom
Residual deviance:   5.656  on  8  degrees of freedom
AIC: 170.07
```

Besides the temperature also the brand M becomes significant and also, to a lesser degree, the previous use of brand M and the combined previous use of brand M plus brand M.

Linear Models


```
Correlation of Coefficients:

(Intercept)                  1
M.userY                      , 1
TempHigh                     , . 1
M.userY:TempHigh             . , , 1
SoftMedium                   , . .   1
SoftSoft                     , . .   . 1
M.userY:SoftMedium           . ,   . , . 1
M.userY:SoftSoft             . ,   . . , . 1
TempHigh:SoftMedium          .   , . . . .   1
TempHigh:SoftSoft            .   , . . .   . . 1
M.userY:TempHigh:SoftMedium  . . . .   , . , . 1
M.userY:TempHigh:SoftSoft    . . .   . . , . , . 1
BrandM                                     .       1
M.userY:BrandM                               .     , 1
TempHigh:BrandM                                  . .   . . 1
M.userY:TempHigh:BrandM                          . .   . . , 1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

Linear Models

compare the two models by an ANOVA table:

```
anova(detg.m0, detg.mod)

Analysis of Deviance Table

Model 1: Fr ~ M.user * Temp * Soft + Brand
Model 2: Fr ~ M.user * Temp * Soft + Brand * M.user * Temp
  Resid. Df Resid. Dev Df Deviance
1        11     32.826
2         8      5.656  3    27.17
```

Tumor Data: Poisson Regression

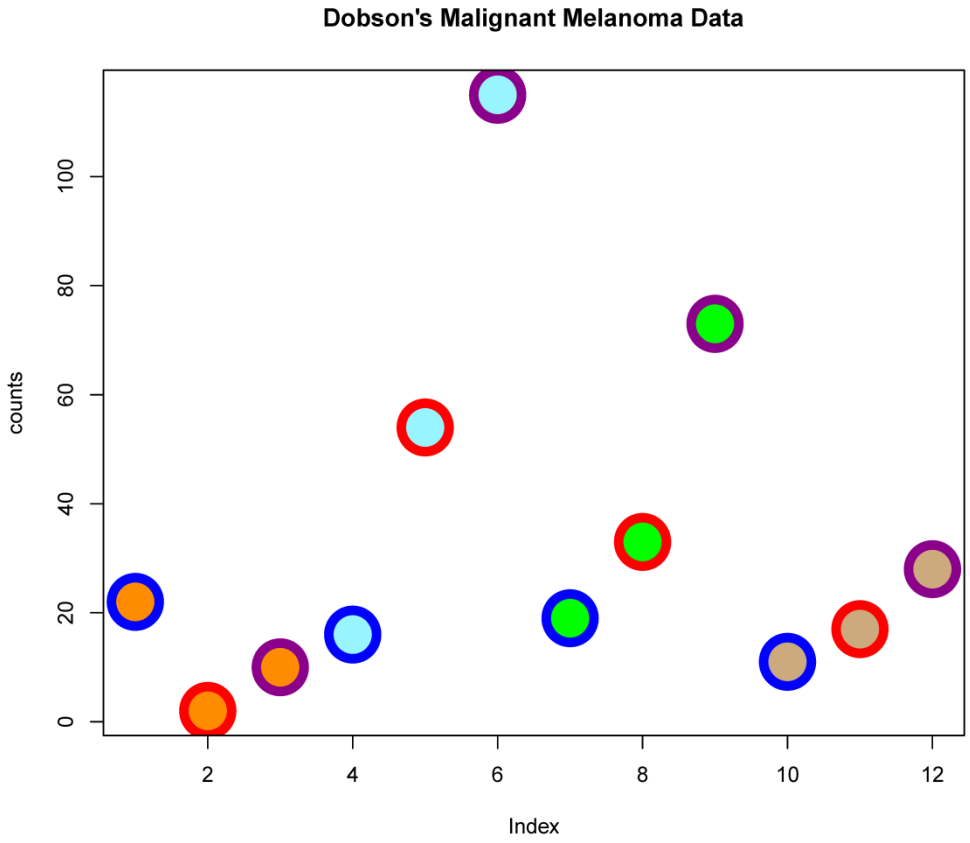
Cross-sectional study of patients with malignant melanoma skin cancer (Roberts 1981). For a sample of $n = 400$ patients, the body site of the tumor and its histological type were determined.

Patients are categorized by type of tumor corresponding to the first factor with four levels: freckle, superficial, nodular, indeterminate. Patients are also categorized by the body site where the tumor was found corresponding to the second factor with three levels: head, trunk, extremities.

Goal: determine association between tumor type and site

	Site			
Tumor type	Head & neck	Trunk	Extrem -ities	Total
Hutchinson's melanotic freckle	22	2	10	34
Superficial spreading melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

Linear Models



The four tumor **types** are indicated by the **interior color** of the circles (orange=freckle, blue=superficial, green=nodular, indeterminate=wood).

The three **locations** at the body are indicated by the **border color** of the circles (head=blue, trunk=red, extremities=magenta).

Linear Models

Poisson regression:

```
summary(z <- glm(counts ~ type + site, family=poisson()))
Call:
glm(formula = counts ~ type + site, family = poisson())
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0453  -1.0741   0.1297   0.5857   5.1354

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.7544     0.2040   8.600  < 2e-16 ***
typesuperficial   1.6940     0.1866   9.079  < 2e-16 ***
typenodular       1.3020     0.1934   6.731 1.68e-11 ***
typeindeterminate 0.4990     0.2174   2.295  0.02173 *
sitetrunk         0.4439     0.1554   2.857  0.00427 **
siteextremities   1.2010     0.1383   8.683  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 295.203  on 11  degrees of freedom
Residual deviance:  51.795  on  6  degrees of freedom
AIC: 122.91
```

Type superficial and nodular are highly significant if compared to type freckle while indeterminate is less significant.

In the figure superficial and nodular have clearly higher counts if compared to freckle. The counts of indeterminate are not so clearly larger.

Site extremities is highly significant. The two largest counts belong to extremities.

To a lesser degree the site trunk is significant. Also seen in the figure, where the third and fourth largest counts with a red border belong to the site trunk.

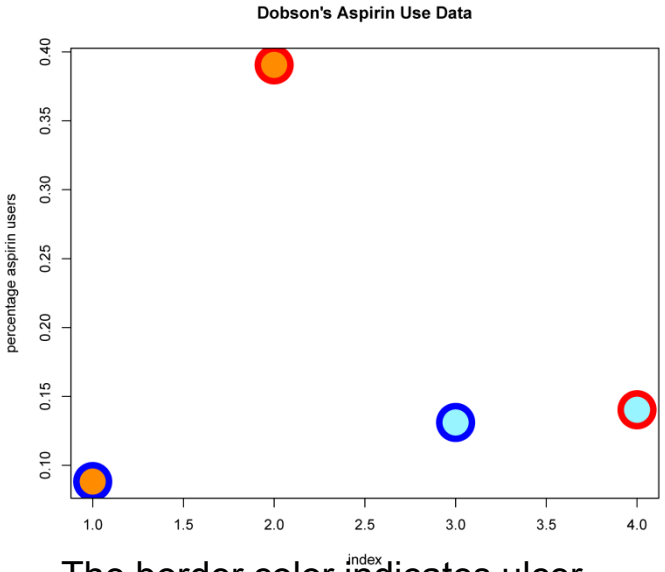
Ulcers and Aspirin Use: Logistic Regression

Retrospective case-control study of gastric and duodenal ulcers and aspirin use (Duggan et al., 1986). Cases and controls are matched with respect to age, sex and socio-economic status.

The individuals are categorized:

1. ulcer cases or controls
2. site of the ulcer: gastric or duodenal
3. aspirin use or not.

	Aspirin use		
	Non-user	User	Total
Gastric ulcer			
Control	62	6	68
Cases	39	25	64
Duodenal ulcer			
Control	53	8	61
Cases	49	8	57
Total	203	47	250



The border color indicates ulcer patients, the cases (red), and controls (blue). The interior color indicates the type of ulcer for the cases: gastric (orange) or duodenal (blue).

Linear Models

Questions of interest:

- Is gastric ulcer associated with aspirin use?
- Is duodenal ulcer associated with aspirin use?
- Is any association with aspirin use the same for both ulcer sites?

model without interaction effects:

```
summary(z<-glm(y~group+type,family=binomial(link="logit")),correlation=TRUE)
Call:
glm(formula = y ~ group + type, family = binomial(link = "logit"))
Deviance Residuals:
      1        2        3        4
 1.2891  -0.9061  -1.5396   1.1959
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.8219     0.3080   5.916  3.3e-09 ***
groupcases    -1.1429     0.3521  -3.246  0.00117 **
typeduodenal   0.7000     0.3460   2.023  0.04306 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 21.789  on 3  degrees of freedom
Residual deviance:  6.283  on 1  degrees of freedom
AIC: 28.003
```

Linear Models


```
Correlation of Coefficients:
             (Intercept) groupcases
groupcases   -0.73
typeduodenal -0.38       -0.05
```

As the count data are not centered, the intercept is significant.

Most significant is the group cases for aspirin use.

The rate is the percentage of the first count of all counts, that is the rate of aspirin non-users. The coefficient of group cases is -1.14 which means the rate of non-users is smaller than the rate for controls.
→ for cases the percentage of aspirin use is larger than for controls.

Less significant and almost not significant is the type of ulcer where gastric is more related to aspirin users.

Linear Models

linear model with interaction effects:

```
summary(z1 <- glm(y ~ group*type, family=binomial(link="logit")))

Call:
glm(formula = y ~ group * type, family = binomial(link = "logit"))

Deviance Residuals:
[1]  0  0  0  0

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.3354     0.4275   5.462  4.7e-08 ***
groupcases         -1.8907     0.4984  -3.793 0.000149 ***
typeduodenal       -0.4445     0.5715  -0.778 0.436711
groupcases:typeduodenal   1.8122     0.7333   2.471 0.013460 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.1789e+01  on 3  degrees of freedom
Residual deviance: 2.3981e-14  on 0  degrees of freedom
AIC: 23.72
```

Again cases are significantly associated with aspirin use. Further cases with gastric are more related to aspirin use.

Linear Models

compare these two models by an ANOVA table:

```
anova(z, z1, test = "Chisq")
Analysis of Deviance Table

Model 1: y ~ group + type
Model 2: y ~ group * type
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         1      6.283
2         0      0.000  1    6.283  0.01219 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The deviance shows that the interaction model is significantly better at fitting the data.

The AIC tells that this may only be due to overfitting to the data:

```
AIC: 28.003          (no interaction)
AIC: 23.72           (interaction)
```

Linear Models

Regularization

In machine learning and statistics it is important to avoid that the model is too much fitted to the data → overfitting.

Overfitting reduces generalization capabilities because other, new data will not have the specific features of the current.

To avoid overfitting, simple models should be selected.

Regularization: To prefer simple models during model

Linear Models

Partial Least Squares Regression

models with $l<m$ variables → regularization by low dimensional space

partial least squares (PLS): factorize both the response Y and X:

$$\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{P}^T + \boldsymbol{E}$$

$$\boldsymbol{Y} = \boldsymbol{U}\,\boldsymbol{Q}^T + \boldsymbol{F}$$

covariance between T and U is maximized
X is an $n \times m$ matrix of predictors
Y is an $n \times p$ matrix of responses
T is an $n \times l$ projection of X
U is an $n \times l$ projection of Y
E is an $n \times m$ additive independently normally distributed noise
F is an $n \times p$ additive independently normally distributed noise

Linear Models

Iterative partial least squares finds projection vectors w for X and v for Y which have maximal covariance:

$$\max_{\|\boldsymbol{w}\|=\|\boldsymbol{v}\|=1} \mathrm{Cov}(\boldsymbol{X}\boldsymbol{w}, \boldsymbol{Y}\boldsymbol{v})$$

Iterative partial least squares is closely related to canonical correlation analysis (CCA) which finds projection vectors w for X and v for Y which have maximal correlation coefficient:

$$\max_{\|\boldsymbol{w}\|=\|\boldsymbol{v}\|=1} \mathrm{corr}(\boldsymbol{X}\boldsymbol{w}, \boldsymbol{Y}\boldsymbol{v})$$

PLS looks at the variance while CCA looks at the correlation.

Linear Models

Partial least squares regression (PLSR) has orthog. score matrix T:

$$\boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I}$$

PLSR defines a linear inner relation (basically a regression):

$$\boldsymbol{U} = \boldsymbol{T}\,\boldsymbol{D} + \boldsymbol{H}$$

D is a diagonal matrix
covariance between T and U is maximized

$$\boldsymbol{Y} = \boldsymbol{T}\,\boldsymbol{D}\,\boldsymbol{Q}^T + \boldsymbol{H}\,\boldsymbol{Q}^T + \boldsymbol{F} = \boldsymbol{T}\,\boldsymbol{C}^T + \boldsymbol{F}'$$

$\boldsymbol{C}^T = \boldsymbol{D}\boldsymbol{Q}^T$: regression coefficients
$\boldsymbol{F}' = \boldsymbol{H}\boldsymbol{Q}^T + \boldsymbol{F}$: noise

least squares estimate with projections T from orthogonal matrices.

Linear Models

noise free case: $\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{P}^T$, $\boldsymbol{T} = \boldsymbol{X}\,\boldsymbol{W}$

$\hat{\boldsymbol{Y}} = \boldsymbol{T}\,\boldsymbol{D}\,\boldsymbol{Q}^T$, $\boldsymbol{U} = \hat{\boldsymbol{Y}}\,\boldsymbol{Q}$

$\hat{\boldsymbol{Y}}$: approximates Y

“latent vectors”: columns of T

“regression weights”: D

“weight matrix” of the dependent variables Y: Q

pseudo inverse: $\boldsymbol{W} = (\boldsymbol{P}^T)^+$

$\boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I}$, $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$, $\boldsymbol{W} = (\boldsymbol{P}^T)^+$

$\boldsymbol{U} = \boldsymbol{T}\,\boldsymbol{D}$, $\boldsymbol{D} = \boldsymbol{T}^T\,\boldsymbol{U}$

Linear Models

Given: matrix $\boldsymbol{X}$, matrix $\boldsymbol{Y}$

initialization

initialize $\boldsymbol{u}$ by random values

$\boldsymbol{A}$ is set to the column centered and column normalized $\boldsymbol{X}$

$\boldsymbol{B}$ is set to the column centered and column normalized $\boldsymbol{Y}$

main loop

while $\boldsymbol{A}$ is not the null matrix **do**

while not converged **do**

$\boldsymbol{w} = \boldsymbol{A}^T\boldsymbol{u}$ (estimate $\boldsymbol{X}$ weights)

$\boldsymbol{t} = \boldsymbol{A}\boldsymbol{w}$ (estimate $\boldsymbol{X}$ factor scores)

$\boldsymbol{t} = \boldsymbol{t}/\|\boldsymbol{t}\|$ (normalize factor scores)

$\boldsymbol{q} = \boldsymbol{B}^T\boldsymbol{t}$ (estimate $\boldsymbol{Y}$ weights)

$\boldsymbol{q} = \boldsymbol{q}/\|\boldsymbol{q}\|$ (normalize weights)

$\boldsymbol{u} = \boldsymbol{B}\boldsymbol{q}$ (estimate $\boldsymbol{Y}$ factor scores)

use $\boldsymbol{w}$ to test if loop has converged

end while

$d = \boldsymbol{t}^T\boldsymbol{u}$

$\boldsymbol{p} = \boldsymbol{A}^T\boldsymbol{t}$

$\boldsymbol{A} = \boldsymbol{A} - \boldsymbol{t}\boldsymbol{p}^T$ (partial out the effect of $\boldsymbol{t}$ from $\boldsymbol{X} \sim \boldsymbol{A}$)

$\boldsymbol{B} = \boldsymbol{B} - d\boldsymbol{t}\boldsymbol{q}^T$ (partial out the effect of $\boldsymbol{t}$ from $\boldsymbol{Y} \sim \boldsymbol{B}$)

store all computed values $\boldsymbol{t}$, $\boldsymbol{u}$, $\boldsymbol{w}$, $\boldsymbol{q}$, $\boldsymbol{p}$ in the corresponding matrices

store d as diagonal element of $\boldsymbol{D}$

end while

result

training: $\hat{\boldsymbol{Y}} = \boldsymbol{T}\boldsymbol{D}\boldsymbol{Q}^T$

prediction: $\boldsymbol{\tau} = \boldsymbol{x}^T\boldsymbol{W}$ ($\boldsymbol{x}$ is normalized like $\boldsymbol{A}$); $\hat{\boldsymbol{y}} = \boldsymbol{\tau}\boldsymbol{D}\boldsymbol{Q}^T$

Partial least squares regression

Linear Models

Partial least squares regression can be based on the singular value decomposition of $\boldsymbol{X}^T\boldsymbol{Y}$

$$\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{P}\,\boldsymbol{T}^T\boldsymbol{U}\,\boldsymbol{Q}^T = \boldsymbol{P}\,\boldsymbol{D}\,\boldsymbol{Q}^T$$

w: largest singular value

q: largest singular value

t: eigenvector of $\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T$

u: eigenvector of $\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{X}^T$

principal components regression: T are the projections onto the first l principal components of X

Ridge Regression

Ridge regression is known as Tikhonov regularization for ill-posed problems.
least squares objective is the sum of squares $\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$

number of regressors is large → overfitting

Regularization by an additional squared term in the parameters:

$$\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \|\boldsymbol{\Gamma}\,\boldsymbol{\beta}\|^2$$

estimator for ridge regression:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Gamma}^T\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Often: $\boldsymbol{\Gamma} = \sqrt{\gamma}\boldsymbol{I}$, where γ is a hyperparameter that controls the trade-off between simple models and low error.

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Linear Models

variance of the ridge regression estimator:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^T\boldsymbol{X} + \gamma\boldsymbol{I}\right)^{-1} \boldsymbol{X}^T\boldsymbol{X} \left(\boldsymbol{X}^T\boldsymbol{X} + \gamma\boldsymbol{I}\right)^{-1}$$

bias of ridge regression estimator:

$$\mathrm{bias}(\hat{\boldsymbol{\beta}}) = -\gamma \left(\boldsymbol{X}^T\boldsymbol{X} + \gamma\boldsymbol{I}\right)^{-1} \boldsymbol{\beta}$$

It has been shown, that there is always a γ for which the parameter mean squared error of ridge regression is smaller than this error of least squares.

Ridge regression is an L^2-norm regularizer, that is the squares of the parameters are weighted and summed up and thereby penalized.

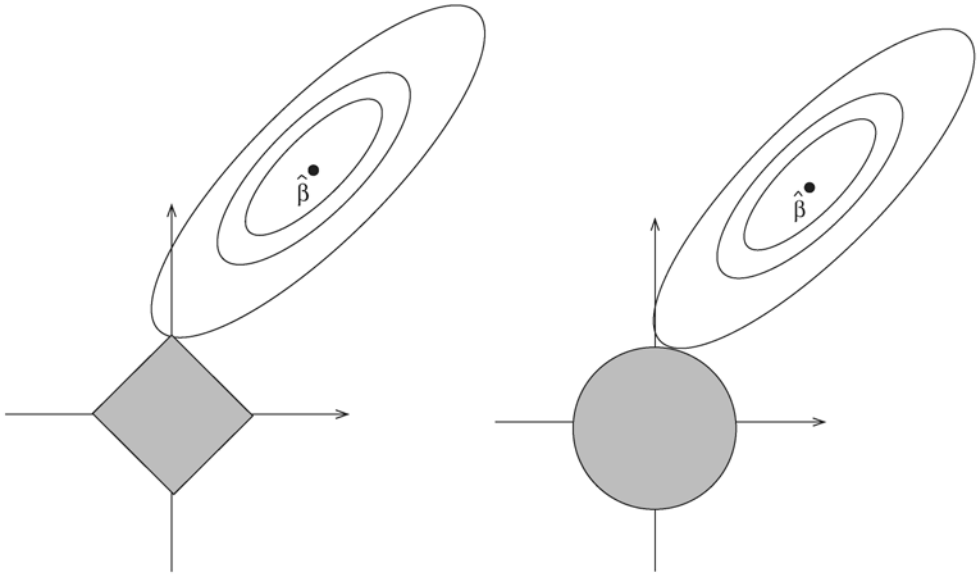
→ small absolute parameter values around zero are preferred
→ But not exactly zero
solution even if the parameters are under-determined because $\left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Gamma}^T\boldsymbol{\Gamma}\right)^{-1}$ always exists → unique solution

Linear Models

Least absolute shrinkage and selection operator (LASSO)

objective $\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \gamma \|\boldsymbol{\beta}\|_1$

In contrast to ridge regression, the LASSO estimate has many zeros



LASSO (left) vs. ridge regression (right). The error objective, the ellipse, touches in most cases a corner of the L^1-norm where at least one component is zero. In contrast the L^2-norm does not possess corners as all points with the same regularization value are on a hyperball.

Linear Models

LASSO is often used for **feature selection** because features, of which the corresponding parameters are zero, can be removed from the model without changing regression result.

The minimization of the LASSO objective is a quadratic optimization problem. It can be solved by techniques of constrained quadratic optimization.

LASSO is consistent if $\gamma/n \xrightarrow{n} 0$

alternative method is the **forward stepwise regression algorithm**:

1. Start with all coefficients β_j equal to zero.
2. Find the predictor x_j which is most correlated with y and add it to the model. Take residuals $r = y - \hat{y}$
3. Continue, at each stage adding to the model the predictor which is most correlated with r.
4. Until: all predictors are in the model

Linear Models

least angle regression procedure is best LASSO estimation method:

- predictor is not fully added only coefficient is increased
- until predictor is no longer correlated with residual r

1. Start with all coefficients β_j equal to zero.
2. Find the predictor x_j that is most correlated with y.
3. Increase the coefficient β_j in the direction of the sign of its correlation with y.
4. Take residuals $r = y - \hat{y}$ and compute correlations.
5. Stop when some other predictor x_k has the same correlation with r than x_j.
6. Increase (β_j, β_k) in their joint least squares direction, until some other predictor x_m has the same correlation with the residual r.
7. Until: all predictors are in the model

This procedure gives the entire path of LASSO solutions if:
a non-zero coefficient is set to zero then remove it from the active set of predictors and recompute the joint direction.

Elastic Net

The L^1-norm has also disadvantages:

- For many features m and few samples n, only the first n features are selected.
- For correlated variables LASSO only selects one variable and does not use the others.

Elastic net is a compromise between ridge regression and LASSO.
It has both an L^1-norm as well as an L^2-norm regularizer.

$$\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \gamma \|\boldsymbol{\beta}\|_1 + \delta \|\boldsymbol{\beta}\|_2^2$$

Problem: two hyperparameters

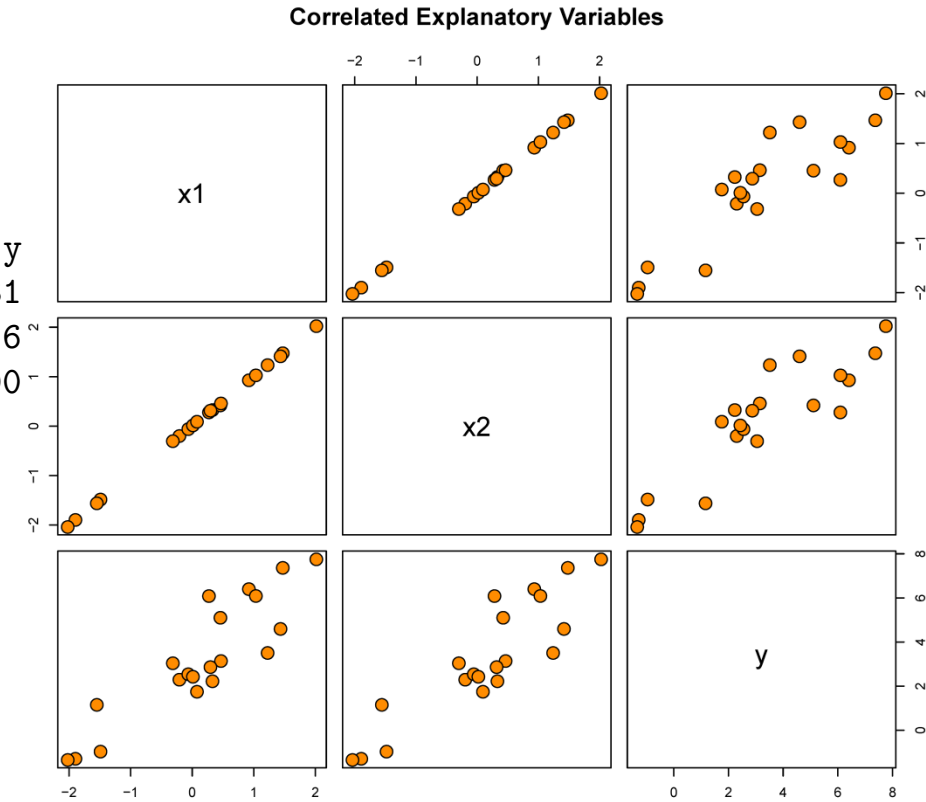
The elastic net is consistent if $\gamma/n \xrightarrow{n} 0$

Examples

Example: Ridge Regression, LASSO, Elastic Net

highly correlated explanatory variables:

```
x1 <- rnorm(20)
x2 <- rnorm(20,mean=x1,sd=.01)
y <- rnorm(20,mean=3+x1+x2)

cor(cbind(x1,x2,y))
          x1        x2        y
x1 1.0000000 0.9999319 0.8927331
x2 0.9999319 1.0000000 0.8919416
y  0.8927331 0.8919416 1.0000000
```



Linear Models

standard linear model:

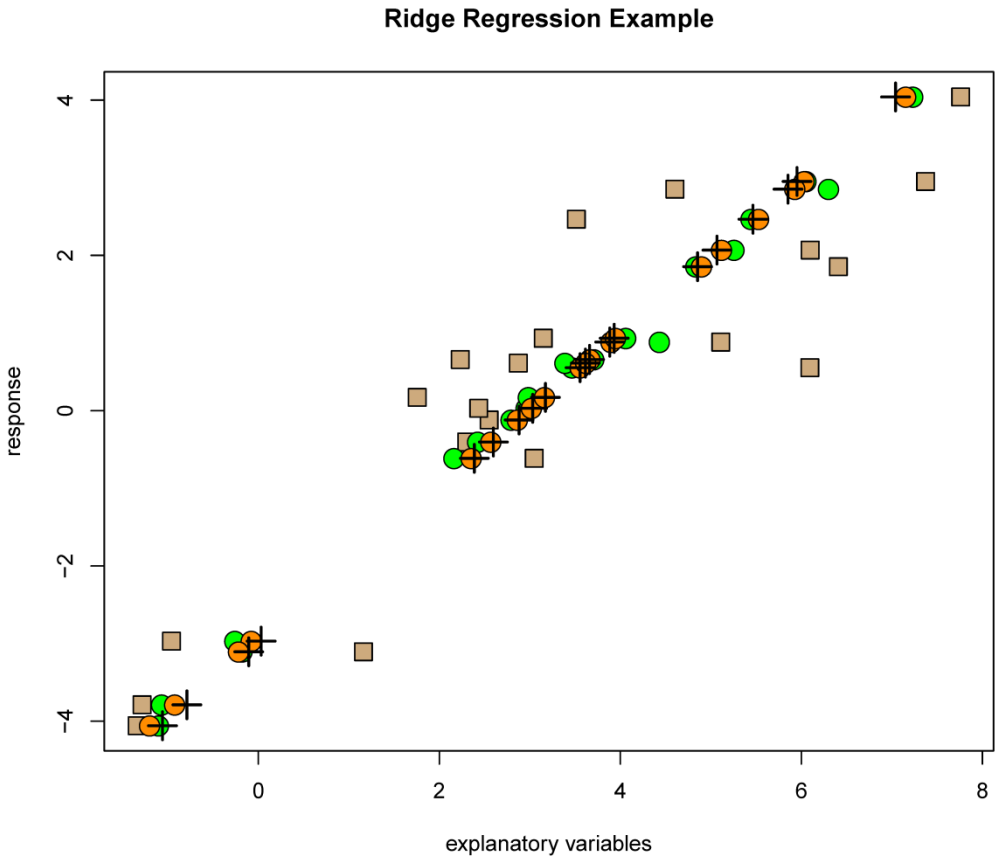
```
l1 <- lm(y~x1+x2)$coef

summary(l1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-12.710  -4.842   3.027   1.723   8.941  14.850

l1
(Intercept)          x1          x2
   3.026583   14.854954  -12.711132
```

Next we fit the model with ridge regression:

```
library(MASS)
l2 <- lm.ridge(y~x1+x2,lambda=1)
l2
                x1        x2
2.985240 1.051382 1.011735
```

Ridge regression is much closer to the true parameter values: 3, 1, 1.

Linear Models



The response data are the wooden-colored squares. Standard least squares gives the green circles while ridge regression gives the orange circles. The noise free data is indicated by crosses. Ridge regression is less prone to overfitting and closer to the crosses and, therefore, it generalizes better.

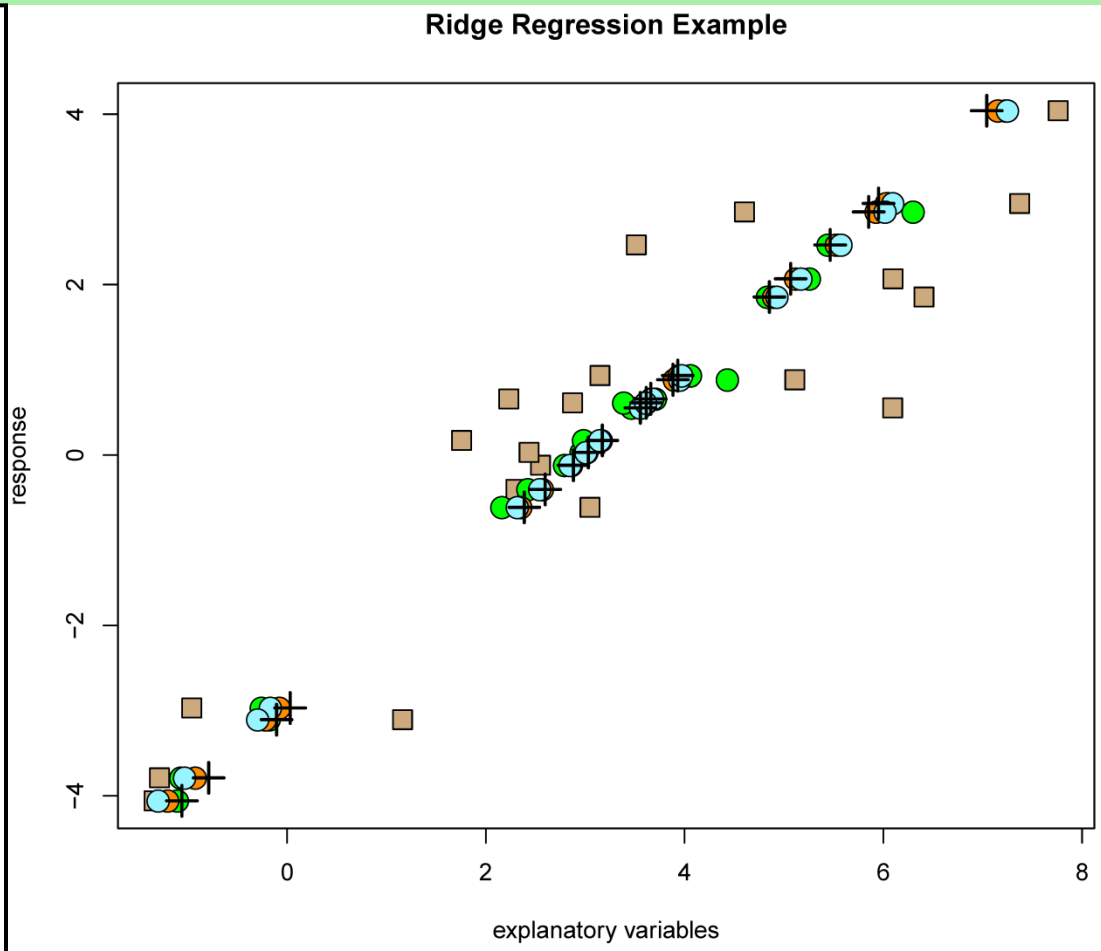
Linear Models


```
library(lars)
l3 <- lars(cbind(x1,x2),y)
Call:
lars(x = cbind(x1, x2), y = y)
R-squared: 0.801
Sequence of LASSO moves:
     x1 x2
Var   1  2
Step  1  2
summary(l3)
LARS/LASSO
Call: lars(x = cbind(x1, x2), y = y)
  Df     Rss      Cp
0  1 138.062 67.3827
1  2  28.030  1.3351
2  3  27.489  3.0000
l3$beta
         x1        x2
0  0.000000   0.00000
1  2.116893   0.00000
2 14.854954 -12.71113
attr(,"scaled:scale")
[1] 4.953151 4.963644
predict(l3,rbind(c(0.0,0.0)))$fit
[1] 3.244128 2.982374 3.026583
```

LASSO solution

The last call supplies the intercepts for the LASSO solutions.
Since in step 2 the residual does not change much compared to step 3 which all variables, we select step 2 solution $y = 2.982374 + 2.116893 x_1$.

Linear Models



LASSO is almost as good as ridge regression since the orange circles are covered by the blue circles obtained from LASSO.
However, LASSO used only one explanatory variable.

Linear Models

elastic net with L^1-norm and L^2-norm equally weighted (α=0.5):

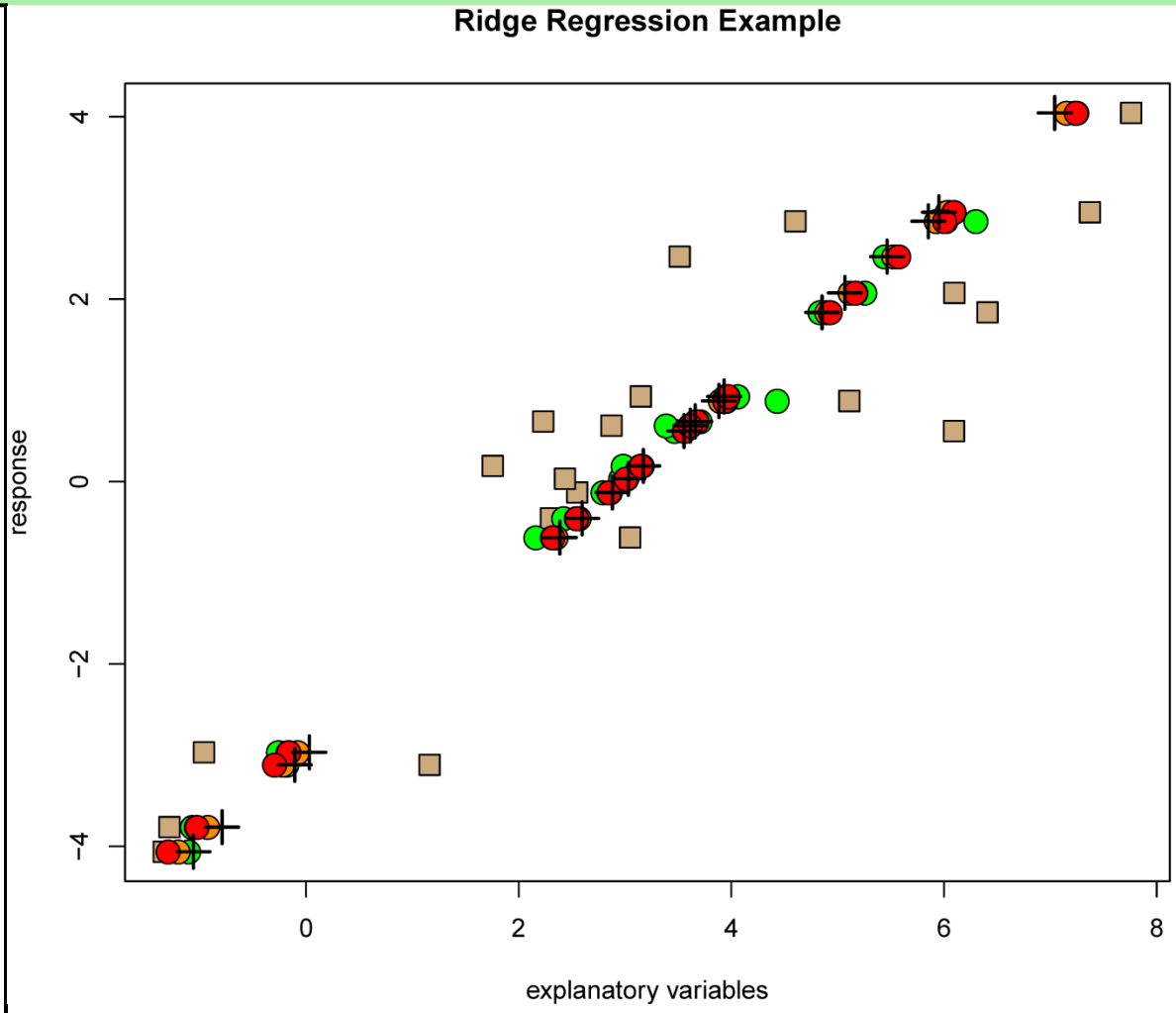
```
library(glmnet)
l4 <- glmnet(cbind(x1,x2),y,alpha=0.5)

summary(l4$lambda)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002078 0.014340 0.098850 0.628400 0.681200 4.691000
```

λ is the factor which weighs the penalty term that includes both L^1- and L^2-norm.

small penalty term:

```
coef(l4,s=0.004)
3 x 1 sparse Matrix of class "dgCMatrix"
    1
(Intercept) 2.981441
x1          1.738632
x2          0.374484
```

Linear Models



The elastic net solution does not differ much from the LASSO solution because the red circles overlay the blue circles.

Linear Models

Example: Diabetes using Least Angle Regression

The data contain blood and other measurements in diabetics and are taken from Efron, Hastie, Johnstone and Tibshirani (2003).
The diabetes data frame has 442 rows and 3 columns:

i. x: a matrix with 10 columns with explanatory variables like age, sex, body mass index (bmi), and blood measurements like cholesterol levels (ldl and hdl) etc.
   1. "age"
   2. "sex2"
   3. "bmi"
   4. "map"
   5. "tc2"
   6. "ldl"
   7. "hdl"
   8. "tch"
   9. "ltg"
   10. "glu"

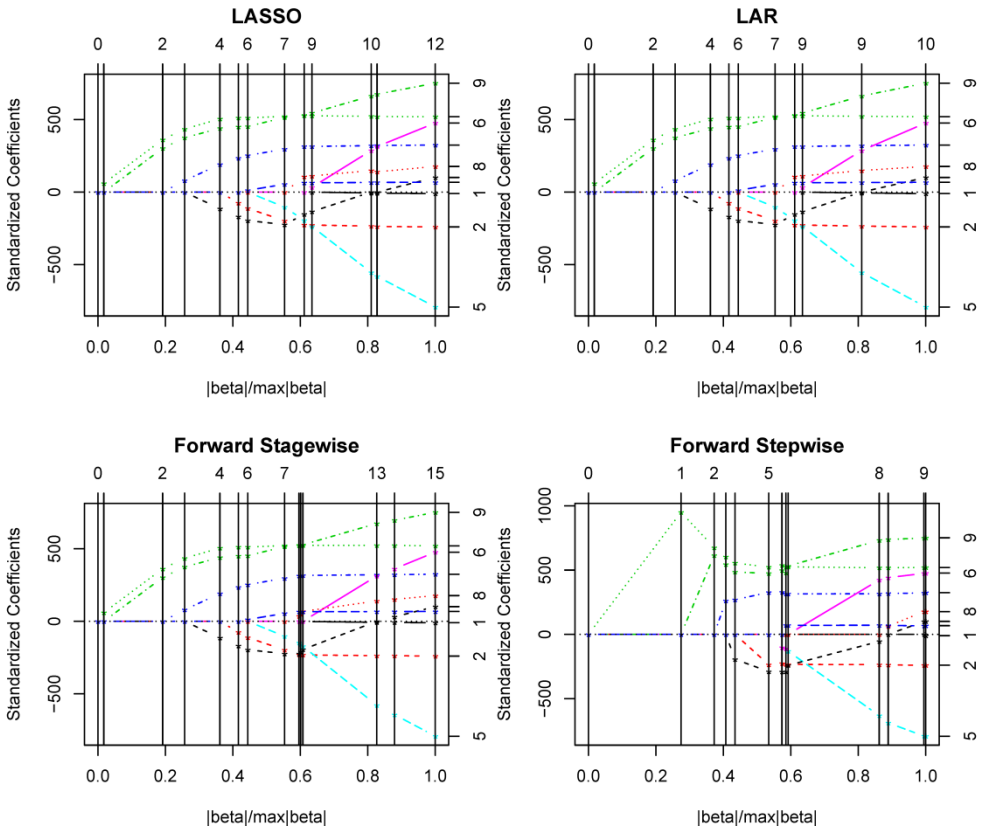
The x matrix has been standardized to have unit L^2-norm in each column and zero mean.
The matrix x_2 consists of x plus certain interactions.

ii. y: a numeric vector

iii. x_2: a matrix with 64 columns of all explanatory variables, their squared values, and measurements of interaction effects.

Linear Models


```
library(lars)
data(diabetes)
x <- diabetes$x
y <- diabetes$y
x2 <- diabetes$x2
op <- par(mfrow=c(2,2))
object1 <- lars(x,y,
+ type="lasso")
plot(object1)
object2 <- lars(x,y,
+ type="lar")
plot(object2)
object3 <- lars(x,y,
+ type="forward.stagewise")
plot(object3)
object4 <- lars(x,y,
+ type="stepwise")
plot(object4)
par(op)
```



The diabetes data set was fitted by LASSO, least angle regression, forward stagewise, and forward stepwise.
The figure shows the coefficients that obtain certain values at certain steps.

Linear Models

solution paths for the different methods:

LASSO

	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu
[1,]	0	0	0	0	0	0	0	0	0	0
[2,]	0	0	60	0	0	0	0	0	0	0
[3,]	0	0	362	0	0	0	0	0	302	0
[4,]	0	0	435	79	0	0	0	0	375	0
[5,]	0	0	506	191	0	0	-114	0	440	0
[6,]	0	-75	511	234	0	0	-170	0	451	0
[7,]	0	-112	512	253	0	0	-196	0	452	12
[8,]	0	-198	522	297	-104	0	-224	0	515	55
[9,]	0	-226	527	314	-195	0	-152	106	530	64
[10,]	0	-227	526	315	-237	34	-135	111	545	65
[11,]	-6	-234	523	320	-554	287	0	149	663	66
[12,]	-7	-237	521	322	-580	314	0	140	675	67
[13,]	-10	-240	520	324	-792	477	101	177	751	68

Linear Models


```
least angle regression
       age  sex  bmi  map   tc  ldl  hdl  tch  ltg  glu
 [1,]    0    0    0    0    0    0    0    0    0    0
 [2,]    0    0   60    0    0    0    0    0    0    0
 [3,]    0    0  362    0    0    0    0    0  302    0
 [4,]    0    0  435   79    0    0    0    0  375    0
 [5,]    0    0  506  191    0    0 -114    0  440    0
 [6,]    0  -75  511  234    0    0 -170    0  451    0
 [7,]    0 -112  512  253    0    0 -196    0  452   12
 [8,]    0 -198  522  297 -104    0 -224    0  515   55
 [9,]    0 -226  527  314 -195    0 -152  106  530   64
[10,]    0 -227  526  315 -237   34 -135  111  545   65
[11,]  -10 -240  520  324 -792  477  101  177  751   68
```

Linear Models


```
forward stagewise
       age  sex  bmi  map   tc  ldl  hdl  tch  ltg  glu
 [1,]    0    0    0    0    0    0    0    0    0    0
 [2,]    0    0   60    0    0    0    0    0    0    0
 [3,]    0    0  362    0    0    0    0    0  302    0
 [4,]    0    0  435   79    0    0    0    0  375    0
 [5,]    0    0  506  191    0    0 -114    0  440    0
 [6,]    0  -75  511  234    0    0 -170    0  451    0
 [7,]    0 -112  512  253    0    0 -196    0  452   12
 [8,]    0 -198  522  297 -104    0 -224    0  515   55
 [9,]    0 -198  522  297 -104    0 -224    0  515   55
[10,]    0 -230  522  313 -148    0 -224   35  524   65
[11,]    0 -231  522  315 -159    0 -211   50  526   66
[12,]    0 -231  522  315 -159    0 -211   50  526   66
[13,]   -1 -232  523  316 -172    0 -195   68  528   66
[14,]   -1 -232  523  316 -172    0 -195   68  528   66
[15,]   -8 -238  523  322 -644  362   31  151  697   67
[16,]  -10 -240  520  324 -792  477  101  177  751   68
```

Linear Models


```
forward stepwise
       age  sex  bmi  map   tc  ldl  hdl  tch  ltg  glu
 [1,]    0    0    0    0    0    0    0    0    0    0
 [2,]    0    0  949    0    0    0    0    0    0    0
 [3,]    0    0  675    0    0    0    0    0  615    0
 [4,]    0    0  603  262    0    0    0    0  544    0
 [5,]    0    0  555  270    0    0 -194    0  485    0
 [6,]    0 -236  524  326    0    0 -289    0  474    0
 [7,]    0 -227  538  328    0 -103 -291    0  498    0
 [8,]    0 -233  527  315    0 -111 -289    0  479   70
 [9,]    0 -236  518  316 -632  423  -55    0  732   71
[10,]    0 -241  520  322 -791  474  100  177  750   66
[11,]  -10 -240  520  324 -792  477  101  177  751   68
```

Linear Models

- final solution is the same
- variables that were selected first and second are the same
- first variable: “bmi” (body mass index), then ”lgt”, “map”, and “hdl”

features selected using data set x_2 (interaction terms):

```
objectN <- lars(x2,y,type="lar")
name <- colnames(x2)
name[which(abs(objectN$beta[2,])>0)]
[1] "bmi"
name[which(abs(objectN$beta[3,])>0)]
[1] "bmi" "ltg"
name[which(abs(objectN$beta[4,])>0)]
[1] "bmi" "map" "ltg"
name[which(abs(objectN$beta[5,])>0)]
[1] "bmi" "map" "hdl" "ltg"
name[which(abs(objectN$beta[6,])>0)]
[1] "bmi"     "map"     "hdl"     "ltg"     "bmi:map"
name[which(abs(objectN$beta[7,])>0)]
[1] "bmi"     "map"     "hdl"     "ltg"     "age:sex" "bmi:map“
```

most important variables are those identified previously

Linear Models

Example: Relevant Variable but No Correlation to Response

toy example where relevant variable is **not correlated to the response**

x_1	-2	2	-2	2
x_2	3	-3	1	-1
t	1	-1	-1	1

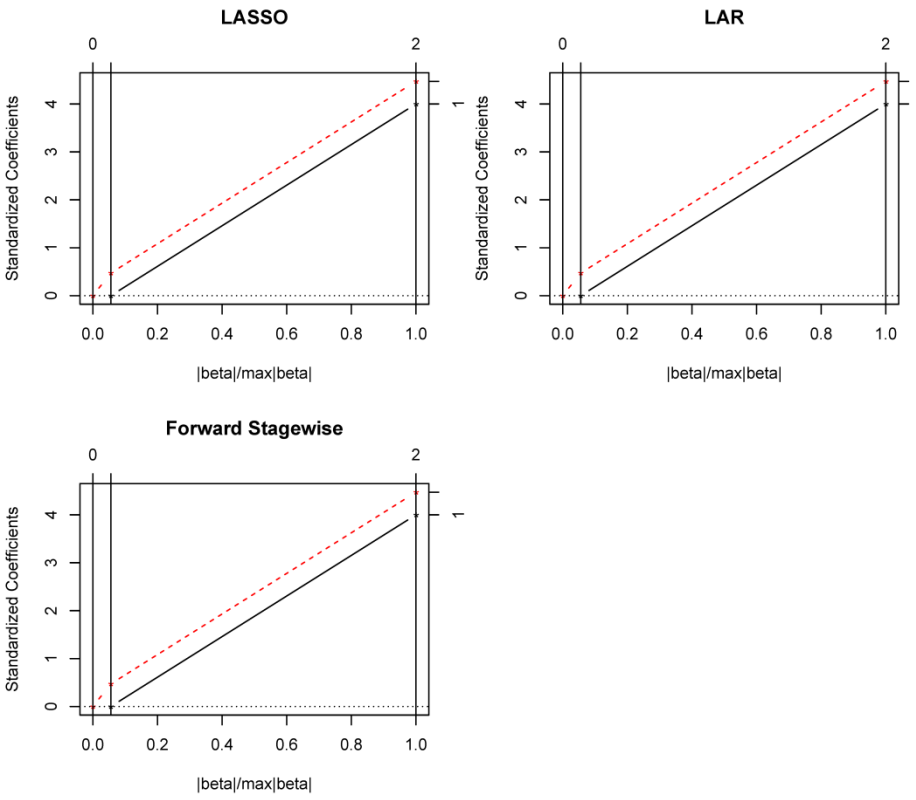
variable x_1 is relevant because $t = x_1 + x_2$ but is not correlated to t

Linear Models

Fitting using least squares regression, ridge regression, LASSO:

```
cor(cbind(t,x1,x2))
           t         x1         x2
t  1.0000000  0.0000000  0.4472136
x1 0.0000000  1.0000000 -0.8944272
x2 0.4472136 -0.8944272  1.0000000

lm(t~x1+x2)$coef
  (Intercept)            x1            x2
-8.326673e-17  1.000000e+00  1.000000e+00

lm.ridge(t~x1+x2,lambda=1)
(Intercept)      x1        x2
0.0000000 0.2622951 0.3278689

e1 <- lars(x,t)
e2 <- lars(x,t,type="lar")
e3 <- lars(x,t,type="for") # Can use abbreviations
op <- par(mfrow=c(2,2))
plot(e1)
plot(e2)
plot(e3)
par(op)
```

Linear Models



solution paths for different LASSO fitting methods

x_1 is always selected in the second step even if it is not correlated to the response variable.

Linear Models

Example: Irrelevant Variable but High Correlation to Response

Toy example with irrelevant variable correlated to the response

x_1	0	1	-1	1
x_2	-1	1	0	0
x_3	0	0	-1	1
t	-1	1	-1	1

variable x_1 is irrelevant because $t = x_2 + x_3$ but is correlated to t

Linear Models


```
cor(cbind(t,x1,x2,x3))
           t        x1        x2        x3
t  1.0000000 0.9045340 0.7071068 0.7071068
x1 0.9045340 1.0000000 0.4264014 0.8528029
x2 0.7071068 0.4264014 1.0000000 0.0000000
x3 0.7071068 0.8528029 0.0000000 1.0000000
```

least squares regression

```
lm(t~x1+x2+x3)$coef
  (Intercept)            x1            x2            x3
-1.171607e-16  4.686428e-16  1.000000e+00  1.000000e+00
```

Least squares finds the correct solution.

ridge regression

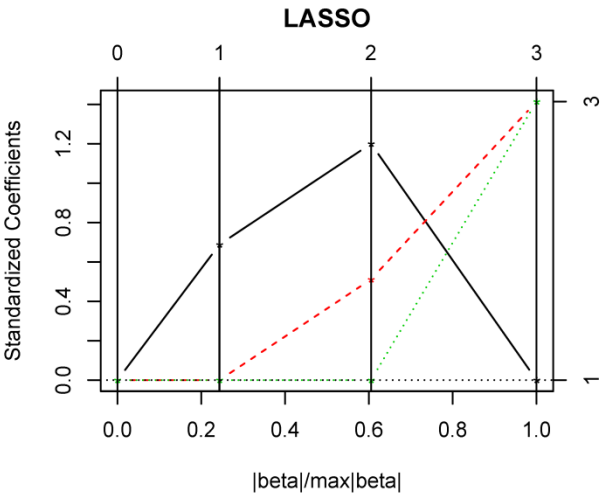
```
lm.ridge(t~x1+x2+x3,lambda=1)
    x1         x2         x3
-0.1043478  0.4173913  0.6330435  0.4660870
```

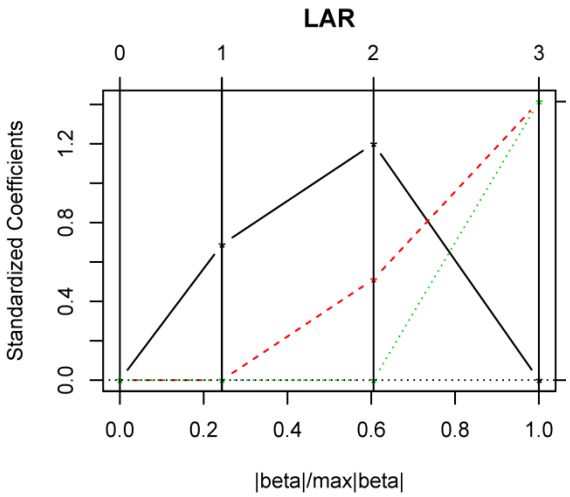
Ridge regression uses the highly correlated variable to reduce the overall squared sum of coefficients (to obtain small regularization terms).

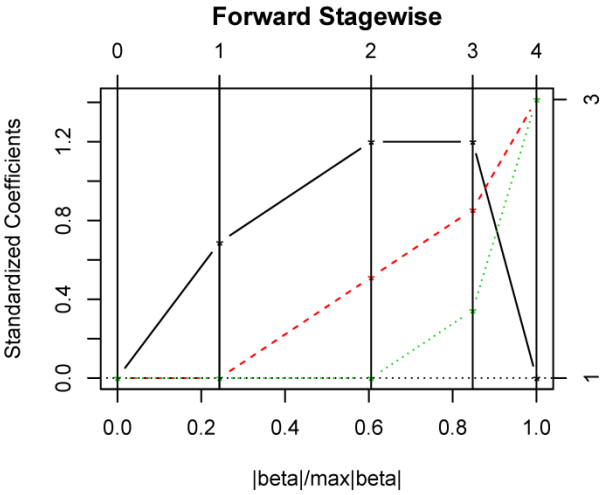
LASSO

```
e1 <- lars(x,t)
e2 <- lars(x,t,type="lar")
e3 <- lars(x,t,type="for") # Can use abbreviations
plot(e1)
plot(e2)
plot(e3)
```

Linear Models







Solution paths for different LASSO fitting methods.

The variable x_1 is selected first but in the last step correctly removed.

Linear Models

Gas Vapor: Ridge Regression and LASSO

When gasoline is pumped into the tank of a car, vapors are vented into the atmosphere (Weisberg, 1985). An experiment was conducted to determine whether the response y, the amount of vapor, can be predicted using the following four variables based on initial conditions of the tank and the dispensed gasoline:

1. x_1: tank temperature (°F),
2. x_2: gasoline temperature (°F),
3. x_3: vapor pressure in tank (psi),
4. x_4: vapor pressure of gasoline (psi).

Linear Models

y	x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4
29	33	53	3.32	3.42	40	90	64	7.32	6.70
24	31	36	3.10	3.26	46	90	60	7.32	7.20
26	33	51	3.18	3.18	55	92	92	7.45	7.45
22	37	51	3.39	3.08	52	91	92	7.27	7.26
27	36	54	3.20	3.41	29	61	62	3.91	4.08
21	35	35	3.03	3.03	22	59	42	3.75	3.45
33	59	56	4.78	4.57	31	88	65	6.48	5.80
34	60	60	4.72	4.72	45	91	89	6.70	6.60
32	59	60	4.60	4.41	37	63	62	4.30	4.30
34	60	60	4.53	4.53	37	60	61	4.02	4.10
20	34	35	2.90	2.95	33	60	62	4.02	3.89
36	60	59	4.40	4.36	27	59	62	3.98	4.02
34	60	62	4.31	4.42	34	59	62	4.39	4.53
23	60	36	4.27	3.94	19	37	35	2.75	2.64
24	62	38	4.41	3.49	16	35	35	2.59	2.59
32	62	61	4.39	4.39	22	37	37	2.73	2.59

gas vapor data from Weisberg (1985)

Linear Models

Correlations

```
cor(m)
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 1.0000000 0.8260665 0.9093507 0.8698845 0.9213333
[2,] 0.8260665 1.0000000 0.7742909 0.9554116 0.9337690
[3,] 0.9093507 0.7742909 1.0000000 0.7815286 0.8374639
[4,] 0.8698845 0.9554116 0.7815286 1.0000000 0.9850748
[5,] 0.9213333 0.9337690 0.8374639 0.9850748 1.0000000
```

The response y is highly correlated with all explanatory variables which in turn are correlated among themselves.
y is most correlated with x_4 followed by x_2.
x_4 is very highly correlated with x_3 and least with x_2.

Linear Models

standard least squares regression:

```
lm(y ~ x)

Call:
l1 <- lm(formula = y ~ x)
l1
Coefficients:
(Intercept)           x1           x2           x3           x4
    1.01502     -0.02861      0.21582     -4.32005      8.97489

anova(l1)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          4 2520.27  630.07   84.54 7.249e-15 ***
Residuals 27  201.23    7.45
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The variables x_3 and x_4 seem to be relevant.
We know that they are highly correlated and lead to overfitting effects.

Linear Models

ridge regression deals with these highly correlated variables:

```
l2 <- lm.ridge(y ~ x,lambda=1)
l2
                      x1           x2           x3           x4
 0.72339986 -0.04937793   0.27780519   0.35225191   3.74029965
```

Here variable x_4 sticks out.

LASSO:

```
la <- lars(x,y,type="lar")
la$beta[2,]
[1] 0.0000000 0.0000000 0.0000000 0.4963341
la$beta[3,]
[1] 0.0000000 0.2695754 0.0000000 3.5437050
la$beta[4,]
[1] -0.06804859  0.27044138  0.00000000  4.48953562
```

Here it becomes clear that x_4 is the most important variable and next the less correlated variable x_2 is selected.

Linear Models

feature selection and use only the variables x_2 and x_4:

```
l3 <- lm(formula = y ~ x[,c(2,4)])
l3

Call:
lm(formula = y ~ x[, c(2, 4)])

Coefficients:
  (Intercept)  x[, c(2, 4)]1  x[, c(2, 4)]2
       0.1918         0.2747         3.6020

anova(l3)
Analysis of Variance Table

Response: y
             Df  Sum Sq Mean Sq F value    Pr(>F)
x[, c(2, 4)]  2 2483.11 1241.56  151.04 4.633e-16 ***
Residuals    29  238.39    8.22
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear Models

We now compare the full model with the model where only two features are selected:

```
anova(l1,l3)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x[, c(2, 4)]
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     27 201.23
2     29 238.39 -2   -37.159 2.4929 0.1015
```

The model with only two features does not perform significantly worse.

check which model is better by Akaike's information criterion (AIC):

```
extractAIC(l1)
[1]  5.00000 68.83842
extractAIC(l3)
[1]  3.00000 70.26103
```

The model with only two variables should be chosen.

Linear Models

Chemical Reaction: Ridge Regression and LASSO

Data set from Box and Youle (1955). The yield in a chemical reaction should be maximized, therefore the values of the following variables were used to control the experiment:

1. x_1: temperature (°C),
2. x_2: concentration of a reagent (%),
3. x_3: time of reaction (hours).

The response variables were:

i. y_1: percent of unchanged starting material,
ii. y_2: percent converted to the desired material.

Linear Models

y_1	y_2	x_1	x_2	x_3
41.5	45.9	162	23	3
33.8	53.3	162	23	8
27.7	57.5	162	30	5
21.7	58.8	162	30	8
19.9	60.6	172	25	5
15.0	58.0	172	25	8
12.2	58.6	172	30	5
4.3	52.4	172	30	8
19.3	56.9	167	27.5	6.5
6.4	55.4	177	27.5	6.5
37.6	46.9	157	27.5	6.5
18.0	57.3	167	32.5	6.5
26.3	55.0	167	22.5	6.5
9.9	58.9	167	27.5	9.5
25.0	50.3	167	27.5	3.5
14.1	61.1	177	20	6.5
15.2	62.9	177	20	6.5
15.9	60.0	160	34	7.5
19.6	60.6	160	34	7.5

Linear Models


```
cor(m)
           [,1]       [,2]       [,3]       [,4]       [,5]
[1,]  1.0000000 -0.60782343 -0.67693865 -0.22472586 -0.45253956
[2,] -0.6078234  1.00000000  0.40395099  0.07998377  0.39273121
[3,] -0.6769387  0.40395099  1.00000000 -0.46200145 -0.02188275
[4,] -0.2247259  0.07998377 -0.46200145  1.00000000  0.17665667
[5,] -0.4525396  0.39273121 -0.02188275  0.17665667  1.00000000
```

y_1: negative correlation to x_1 and less negative correlation to x_3

y_2: negatively correlated to y_1 (expected) and equally correlated to x_1 and x_3

Linear Models


```
l1 <- lm(y1 ~ x)
l1

Call:
lm(formula = y1 ~ x)

Coefficients:
(Intercept)           x1           x2           x3
    332.111       -1.546       -1.425       -2.237

anova(l1)
Analysis of Variance Table

Response: y1
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          3 1707.16  569.05  106.47 2.459e-10 ***
Residuals 15   80.17    5.34
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

All variables are relevant for prediction.
x_3 is the most relevant variable.

Linear Models

regularization using ridge regression:

```
l2 <- lm.ridge(y1 ~ x,lambda=1)
l2
    x1          x2          x3
307.512361  -1.424838  -1.279060  -2.179261
```

no change compared to standard least squares → all variables are required.

LASSO:

```
la <- lars(x,y1,type="lar")
la$beta[2,]
[1] -0.3518723  0.0000000  0.0000000
la$beta[3,]
[1] -0.5182233  0.0000000 -0.6334936
```

The first and last variable seem to be the most relevant ones.

Linear Models

least squares model with the two most important variables:

```
l3 <- lm(formula = y1 ~ x[,c(1,3)])
l3

Call:
lm(formula = y1 ~ x[, c(1, 3)])

Coefficients:
  (Intercept)  x[, c(1, 3)]1  x[, c(1, 3)]2
      222.957         -1.101         -2.853


anova(l3)
Analysis of Variance Table

Response: y1
             Df  Sum Sq Mean Sq F value    Pr(>F)
x[, c(1, 3)]  2 1209.61  604.81   16.75 0.0001192 ***
Residuals    16  577.72   36.11
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear Models

ANOVA table shows that all variables are required for prediction:

```
anova(l1,l3)
Analysis of Variance Table

Model 1: y1 ~ x
Model 2: y1 ~ x[, c(1, 3)]
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     15  80.17
2     16 577.72 -1   -497.55 93.088 7.988e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear Models

second response variable y_2: converted material to the desired product

```
l12 <- lm(y2 ~ x)
l12

Call:
lm(formula = y2 ~ x)

Coefficients:
(Intercept)           x1           x2           x3
   -26.0353       0.4046       0.2930       1.0338

anova(l12)
Analysis of Variance Table

Response: y2
          Df Sum Sq Mean Sq F value  Pr(>F)
x          3 151.00  50.334  3.0266 0.06235 .
Residuals 15 249.46  16.631
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Again x_3 is the most relevant variable - now even more dominant

Linear Models

ridge regression:

```
l22 <- lm.ridge(y2 ~ x,lambda=1)
l22
      x1          x2          x3
-19.9403245   0.3747668   0.2617700   0.9933463
```

The figure remains the same

We perform fitting with LASSO:

```
la2 <- lars(x,y2,type="lar")
la2$beta[2,]
[1] 0.008327752 0.000000000 0.000000000
la2$beta[3,]
[1] 0.1931751 0.0000000 0.7039310
```

Interestingly, x_1 is selected before x_3
Looking at the correlation matrix, we see that indeed x_1 is more correlated to y_2 than x_3 (0.40 vs. 0.39).

Linear Models

least squares on two variables first selected by LASSO:

```
l32 <- lm(formula = y2 ~ x[,c(1,3)])
l32

Call:
lm(formula = y2 ~ x[, c(1, 3)])

Coefficients:
  (Intercept)  x[, c(1, 3)]1  x[, c(1, 3)]2
      -3.5856         0.3131         1.1605

anova(l32)
Analysis of Variance Table

Response: y2
             Df Sum Sq Mean Sq F value  Pr(>F)
x[, c(1, 3)]  2 129.96  64.978  3.8433 0.04334 *
Residuals    16 270.51  16.907
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear Models

Comparing the full model with this model by an ANOVA table:

```
anova(l12,l32)
Analysis of Variance Table

Model 1: y2 ~ x
Model 2: y2 ~ x[, c(1, 3)]
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     15 249.46
2     16 270.51 -1   -21.047 1.2655 0.2783
```

→ the model with only two features is not significantly worse than the full model.

Linear Models

Land Rent: Ridge Regression and LASSO

For 34 counties in Minnesota the following variables were recorded in 1977 (Weisberg, 1985):

1. y: average rent paid per acre of land with alfalfa,
2. x_1: average rent paid per acre for all land,
3. x_2: average number of dairy cows per square mile,
4. x_3: proportion of farmland in pasture.

A relevant question is:
can the rent for alfalfa land be predicted from the other three variables?

Linear Models

y	x_1	x_2	x_3	y	x_1	x_2	x_3
18.38	15.50	17.25	.24	8.50	9.00	8.89	.08
20.00	22.29	18.51	.20	36.50	20.64	23.81	.24
11.50	12.36	11.13	.12	60.00	81.40	4.54	.05
25.00	31.84	5.54	.12	16.25	18.92	29.62	.72
52.50	83.90	5.44	.04	50.00	50.32	21.36	.19
82.50	72.25	20.37	.05	11.50	21.33	1.53	.10
25.00	27.14	31.20	.27	35.00	46.85	5.42	.08
30.67	40.41	4.29	.10	75.00	65.94	22.10	.09
12.00	12.42	8.69	.41	31.56	38.68	14.55	.17
61.25	69.42	6.63	.04	48.50	51.19	7.59	.13
60.00	48.46	27.40	.12	77.50	59.42	49.86	.13
57.50	69.00	31.23	.08	21.67	24.64	11.46	.21
31.00	26.09	28.50	.21	19.75	26.94	2.48	.10
60.00	62.83	29.98	.17	56.00	46.20	31.62	.26
72.50	77.06	13.59	.05	25.00	26.86	53.73	.43
60.33	58.83	45.46	.16	40.00	20.00	40.18	.56
49.75	59.48	35.90	.32	56.67	62.52	15.89	.05

Linear Models

correlation:

```
cor(m)
           [,1]       [,2]      [,3]       [,4]
[1,]  1.0000000  0.8868392 0.2967901 -0.3838808
[2,]  0.8868392  1.0000000 0.0296753 -0.5212982
[3,]  0.2967901  0.0296753 1.0000000  0.4876448
[4,] -0.3838808 -0.5212982 0.4876448  1.0000000

sd(m[,1])
[1] 21.53698
sd(m[,2])
[1] 22.45614
sd(m[,3])
[1] 14.21056
sd(m[,4])
[1] 0.1532131
```

We computed the standard deviations of the variables because x_3 has smaller values than the other variables.

x_3 is about a factor of 100 smaller than the other variables.

Linear Models

least squares regression:

```
l1 <- lm(y ~ x)
l1

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)           x1           x2           x3
     0.6628       0.7803       0.5031     -17.1002

anova(l1)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          3 13266.9  4422.3  65.037 3.112e-13 ***
Residuals 30  2039.9    68.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

x_3 has the largest coefficient but it has to be divided by a factor of 100 to be in the range of the other variables.

→ x_3 has actually the smallest influence on the response

Linear Models

Ridge regression:

```
l2 <- lm.ridge(y ~ x,lambda=1)
l2
                     x1          x2          x3
  2.1360609   0.7542789   0.4955992 -18.2104311
```

Ridge regression penalizes the coefficients for the standardized variables, the absolute coefficient for x_3 even increases.
The other two coefficients decrease as they are pushed toward zero.

LASSO:

```
la <- lars(x,y,type="lar")
la$beta[2,]
[1] 0.5832042 0.0000000 0.0000000
la$beta[3,]
[1] 0.7872064 0.3223731 0.0000000
```

x_1 and x_2 are the most relevant as expected from the correlations, where x_1 has largest correlation with the response.
Interestingly, x_3 has the second largest correlation to the response but is not selected: x_3 has also large correlation to x_1 and does not bring in much new information.
In contrast x_2 has low correlation to x_1 and brings in new information.

Linear Models

least squares for the first two explanatory variables:

```
l3 <- lm(formula = y ~ x[,c(1,2)])
l3

Call:
lm(formula = y ~ x[, c(1, 2)])

Coefficients:
  (Intercept)  x[, c(1, 2)]1  x[, c(1, 2)]2
      -3.3151         0.8428         0.4103

anova(l3)
Analysis of Variance Table

Response: y
             Df  Sum Sq Mean Sq F value    Pr(>F)
x[, c(1, 2)]  2 13159.3  6579.6  94.981 6.015e-14 ***
Residuals    31  2147.5    69.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear Models

Comparing the full model with the model that has only the first two variables shows that the error difference is not significant:

```
anova(l1,l3)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x[, c(1, 2)]
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     30 2039.9
2     31 2147.5 -1   -107.58 1.5821 0.2182
```

Therefore the reduce model may be chosen for analysis.