Existence of the MLE

Corollary (Cor. 6.11)

The set $C = \tau(\mathbb{R}^p)$ has the representation

$$C = \left\{ \sum_{i=1}^{n} \mu_i X_i \mid \mu_i \in J \right\} \tag{1}$$

and is convex. If $\mathbf{X}$ has full rank p then C is open.

To check if the MLE exists we need to check if $t \in C$. This is trivially the case with probability 1 if

$$P(Y \in J) = 1$$

but less trivial to check if $P(Y \in \partial J) > 0$.

The solution, if it exists, is unique if $\mathbf{X}$ has full rank p.

Poisson example

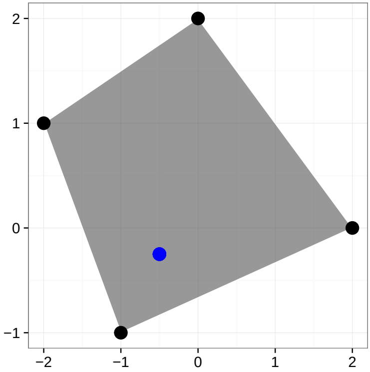
```
X <- data.frame(x1 = c(-2, -1, 2, 0), x2 = c(1, -1, 0, 2))
y <- c(1, 2, 1, 0); Xy <- cbind(y, X)
t <- c(y %*% X$x1, y %*% X$x2) / sum(y)
summary(glm(y ~ x1 + x2, family = poisson, data = Xy))
```

```
...
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.0267     0.5461    0.05     0.96
## x1           -0.1237     0.3742   -0.33     0.74
## x2           -0.6550     0.5157   -1.27     0.20
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2.77259  on 3  degrees of freedom
## Residual deviance: 0.75402  on 1  degrees of freedom
## AIC: 13.37
##
## Number of Fisher Scoring iterations: 5
```

Poisson example

In the example the average t was in the interior of the convex hull, and we could fit the Poisson model using `glm`.

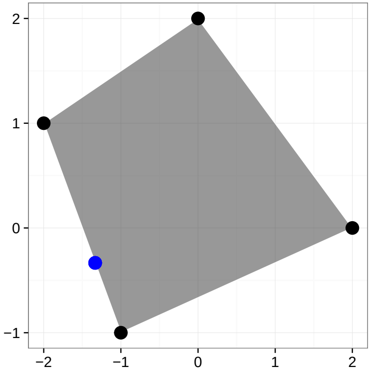
```
p <- qplot(x1, x2, data = X, geom = "polygon", alpha = I(0.5)) +
  geom_point(size = 5, alpha = 1) + xlab("") + ylab("")
p + geom_point(aes(t[1], t[2]), size = 5, color = "blue")
```



Poisson example

Then we consider an example where the average t ends up on the boundary of the convex hull.

```
y <- c(1, 2, 0, 0); Xy <- cbind(y, X)
t <- c(y %*% X$x1, y %*% X$x2) / sum(y)
p + geom_point(aes(t[1], t[2]), size = 5, color = "blue")
```



Poisson example

```
summary(glm(y ~ x1 + x2, family = poisson, data = Xy))


...
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -12.84   25534.58       0        1
## x1             -8.79   17023.05       0        1
## x2             -4.74    8511.53       0        1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4.4987e+00  on 3  degrees of freedom
## Residual deviance: 4.0610e-10  on 1  degrees of freedom
## AIC: 10.61
##
## Number of Fisher Scoring iterations: 21
```

Binary response

If the response is binary, $I = \mathbb{R}$, $J = (0,1)$ and the canonical link is the logit function

$$(0,1) \ni p \mapsto \mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

The response variables all take values on the boundary of $J = (0,1)$!

We have that

$$t = \sum_{i:Y_i=1} X_i \in \overline{C}$$

and we need to find conditions in terms of the X_i that ensure that $t \in C$.

Separation

The responses $Y_1, \ldots, Y_n \in \{0, 1\}$ are binary.

Definition

We say that $X_1, \ldots, X_n \in \mathbb{R}^p$ are separated by $Y_1, \ldots, Y_n$ if there exists a nonzero vector $\beta \in \mathbb{R}^p$ such that for all $i = 1, \ldots, n$

$$X_i^T \beta \geq 0 \quad \text{if } Y_i = 1,$$

and

$$X_i^T \beta \leq 0 \quad \text{if } Y_i = 0.$$

Observe that if $\mathbf{X}$ has full rank p, and the rows are separated according to the definition above, the at least one of the n inequalities above is sharp because β is assumed nonzero. The vector β is called the separating vector.

Existence of the MLE in logistic regression

We consider binary responses $Y_1, \ldots, Y_n \in \{0, 1\}$ and the logistic regression model.

Theorem (Th, 6.16)

Assume that $\mathbf{X}$ *has full rank* p*. The MLE exists if and only if the rows of* $\mathbf{X}$ *are not separated by* $Y_1, \ldots, Y_n$*.*

Being explicit about the intercept

If the model contains an intercept in addition to the predictors $X_i \in \mathbb{R}^p$, it is

$$\tilde{X}_i = (1, X_i^T)^T$$

for $i = 1, \ldots, n$ that must be checked for separability. This is equivalent to the existence of $\beta \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}$ such that for all $i = 1, \ldots, n$

$$X_i^T \beta \geq \beta_0 \quad \text{if } Y_i = 1,$$

and

$$X_i^T \beta \leq \beta_0 \quad \text{if } Y_i = 0.$$

Checking for linear separability

Corollary (Cor. 6.17)

Assume that $\mathbf{X}$ *has full rank* p. *The maximization problem*

$$\begin{array}{ll} \textit{maximize} & \sum_{i=1}^{n} s_i \\ \textit{subject to} & (2Y_i - 1)X_i^T \beta \geq s_i, \ s_i \geq 0, \ i = 1, \ldots, n, \\ & -1 \leq \beta_j \leq 1, \ j = 1, \ldots, p \end{array}$$

in the variables $(\beta^T, s^T)^T \in \mathbb{R}^{n+p}$ *has a solution with* $\sum_{i=1}^{n} s_i > 0$ *if and only if* $X_1, \ldots, X_n$ *are separated by* $Y_1, \ldots, Y_n$.

The constraints on the β_j's force the s_i's to be bounded, and the constraints are fulfilled for $\beta = \mathbf{0}_p$ and $s = \mathbf{0}_n$. Thus we maximize a linear function over a compact set, and there is always a finite solution bounded below by 0.

Poisson responses

For Poisson distributed responses we have $I = \mathbb{R}$, $J = (0, \infty)$ and canonical link

$$(0, \infty) \ni \mu \mapsto \log(\mu).$$

The nonexistence of the MLE is clearly related to observations being 0.

Existence of the MLE in Poisson regression

We consider positive responses $Y_i \geq 0$ and the Poisson regression model with log-link. We let

$$t_0 = \sum_{i=1}^{n} Y_i X_i = \mathbf{X}^T \mathbf{Y}.$$

Corollary (Cor. 6.13)

Assume that $\mathbf{X}$ *has full rank p. The MLE exists if and only if the following linear program*

$$\begin{array}{ll} \textit{maximize} & s \\ \textit{subject to} & \mathbf{X}^T \mu = t_0,\ \mu_i - s \geq 0,\ s \geq 0. \end{array}$$

in the variables $(\mu^T, s)^T \in \mathbb{R}^{n+1}$ *has a feasible point with* $s > 0$*.*

Note that $(\mathbf{Y}^T, 0)^T$ is a feasible point.

Specifying the linear program in practice

The linear program is specified in practice in terms of a vector $c \in \mathbb{R}^{n+1}$ of objective coefficients and an $(n+p) \times (n+1)$ constraint matrix A. They are given as

$$c = (0, \ldots, 0, 1)^T$$

and

$$A = \left(\begin{array}{cc} \mathbf{I}_n & -\mathbf{1}_n \\ \mathbf{X}^T & \mathbf{0}_p \end{array} \right)$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix, $\mathbf{1}_n$ is the n-dimensional vector of ones and $\mathbf{0}_p$ is the p-dimensional vector of zeroes.

The constraint matrix specifies the left hand side of the $n+p$ constraints in the $n+1$ variables. The first n are inequality constraints and the last p are equality constraints. The right hand side of the constraints is the $(n+p)$-dimensional vector

$$\left(\begin{array}{c} \mathbf{0}_n \\ t_0 \end{array} \right).$$

Poisson example

```
## Coefficient vector
c <- c(0, 0, 0, 0, 1)
## Constraint matrix
A <- matrix(
  c(1, 0, 0, 0, -1,
    0, 1, 0, 0, -1,
    0, 0, 1, 0, -1,
    0, 0, 0, 1, -1,
    1, 1, 1, 1,  0,
   -2, -1, 2, 0, 0,
    1, -1, 0, 2, 0),
  nrow = 7,
  ncol = 5,
  byrow = TRUE)
## Right hand side
t <- A[5:7, 1:4] %*% c(1, 2, 1, 0)
rhs <- c(0, 0, 0, 0, t)
## Directions of the (in)equalities
dir <- c(rep(">=", 4), rep("=", 3))
```

Poisson example

```r
lp(direction = "max",
   objective.in = c,
   const.mat = A,
   const.dir = dir,
   const.rhs = rhs
)

## Success: the objective function is 0.47
```

Poisson example

```
## Changing the right hand side
t <- A[5:7, 1:4] %*% c(1, 2, 0, 0)
rhs <- c(0, 0, 0, 0, t)
## Solving the linear program
lp(direction = "max",
   objective.in = c,
   const.mat = A,
   const.dir = dir,
   const.rhs = rhs
)

## Success: the objective function is 0
```

The idealized estimator

Under GA1, GA2 and A4, and with $\mathbf{Z}$ and $\mathbf{W}$ the working response and weight matrix in the true β, then with

$$\hat{\beta}^{\text{ideal}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Z}$$

we have

$$\begin{aligned} E(\hat{\beta}^{\text{ideal}} \mid \mathbf{X}) &= \beta, \\ V(\hat{\beta}^{\text{ideal}} \mid \mathbf{X}) &= \psi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}, \\ E(||\mathbf{Z} - \mathbf{X}\hat{\beta}^{\text{ideal}}||^2_{\mathbf{W}} \mid \mathbf{X}) &= (n-p)\psi. \end{aligned}$$

Deviance

Let $\hat{\mu}_i$ be the MLE of μ_i in a glm based on response observations $Y_1, \ldots, Y_n$.

Definition

The deviance is

$$D = \sum_{i=1}^{n} d(Y_i, \hat{\mu}_i).$$

For a linear hypothesis H_0 on the β-parameter the corresponding deviance is

$$D_0 = \sum_{i=1}^{n} d(Y_i, \hat{\mu}_i^0),$$

the deviance test statistic is $D_0 - D$ and the F-test statistic is

$$\frac{(D_0 - D)/(p - p_0)}{D/(n-p)}.$$

Example

```r
x1 <- rnorm(100); x2 <- factor(rbinom(100, 2, 0.2))
beta <- c(0.3, 0.4, 0.6, 0.1)
y <- rpois(100, exp(beta[1] * x1 + beta[as.numeric(x2) + 1]))
simGlm <- glm(y ~ x1 + x2, family = "poisson")
summary(simGlm)


...
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.4001     0.0990    4.04  5.3e-05
## x1            0.2770     0.0903    3.07   0.0022
## x21           0.2220     0.1692    1.31   0.1896
## x22          -1.8015     1.0048   -1.79   0.0730
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 134.79  on 99  degrees of freedom
## Residual deviance: 115.60  on 96  degrees of freedom
## AIC: 313.5
##
## Number of Fisher Scoring iterations: 5
```

Example

```
anova(simGlm, test = "Chisq")

...
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                    99        135
## x1    1    10.45        98        124   0.0012
## x2    2     8.74        96        116   0.0127
```

```
drop1(simGlm, test = "Chisq")

...
##        Df Deviance AIC  LRT Pr(>Chi)
## <none>          116 314
## x1      1       125 321 9.68   0.0019
## x2      2       124 318 8.74   0.0127
```

Residuals

Raw residuals

$$y_i - \hat{\mu}_i.$$

Pearson residuals

$$\frac{Y_i - \hat{\mu}_i}{\sqrt{\mathcal{V}(\hat{\mu}_i)}}.$$

Deviance residuals

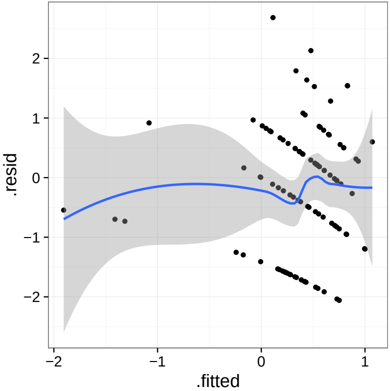
$$\mathrm{sign}(Y_i - \hat{\mu}_i)\sqrt{d(Y_i, \hat{\mu}_i)}.$$

Working residuals

$$\frac{Y_i - \hat{\mu}_i}{\mu'(\hat{\eta}_i)}.$$

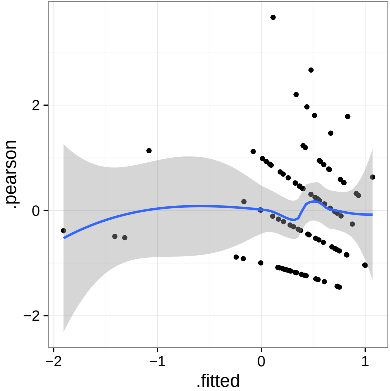
Deviance residuals

```
simDiag <- fortify(simGlm)
qplot(.fitted, .resid, data = simDiag) +
  geom_smooth()
```



Example

```
simDiag$.pearson <- residuals(simGlm, type = "pearson")
qplot(.fitted, .pearson, data = simDiag) +
  geom_smooth()
```



Model validation

For the linear model an index of predictive ability is R^2 – or rather adjusted $\overline{R}^2$, which is not too optimistic for complex models.

For generalized linear models we can replace RSS in the definition of R^2 or $\overline{R}^2$ by

$$\mathcal{X}^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\mathcal{V}(\hat{\mu}_i)},$$

the Pearson χ^2-statistic, or by the deviance

$$D = \sum_{i=1}^{n} d(Y_i, \hat{\mu}_i).$$

The numerical value of these pseudo-R^2 statistics are not comparable to what is obtained for the linear model.

Model validation and selection

We estimate the dispersion parameter as

$$\hat{\psi} = \frac{1}{n-p}\mathcal{X}^2(p),$$

For selection of a submodel of dimension p_0 one can minimize a pseudo-R^2 or the AIC, which (for fixed dispersion parameter) is*

$$\mathrm{AIC} = D(p_0)/\psi + 2p_0.$$

Or Mallows's C_p statistic

$$C_p = D(p_0) + 2\hat{\psi}p_0.$$

* This may only equal what R produces up to an additive constant, and there are additional nuances when ψ is estimated.

Using C_p for model comparison

```
anova(simGlm, test = "Cp")


...
##      Df Deviance Resid. Df Resid. Dev  Cp
## NULL                    99        135 137
## x1    1    10.45        98        124 128
## x2    2     8.74        96        116 124
```

AIC, C_p or (pseudo-) R^2 quantify predictive strength of the model on data – predictors and responses – sampled from the same distribution as the data used to fit the model (Chapter 8, not in course).

They do not quantify model fit!

Model selection consequences

Classical sampling properties of estimators and test statistics are invalidated by model selection.

Sampling distributions of the combined procedure

$$\text{model selection} + \text{parameter estimation}$$

are nonstandard, difficult to derive in theory, and prone to change fundamentally depending on the model selection method.

Bootstrapping can (partially) alleviate the problem.