第七章 伪似然估计及其统计推断

第七章 伪似然估计及其统计推断

Quasi-Likelihood: Motivation

Very often it is hard to specify a pdf or pmf for our data of interest.

Instead we may be willing to specify:

1. Whether or not the random variables are independent.
2. The mean of the random variables.
3. The variance of the random variables

Can we do inference in this case?

GLM setup

For a GLM we might suppose that

1. $\{Y_1, ..., Y_n\}$ are a set of independent random variables.

2. $\mu_i = E(Y_i)$ is related to a set of covariates x_i and some parameter vector β through some link function $g(\cdot)$.

3. $var(Y_i) = \sigma^2 V_i(\mu_i)$ where σ^2 may be unknown and $V_i(\mu_i)$ models the mean variance relationship for each i.

(Commonly we assume that $V_i(\mu_i) = V(\mu_i)$ for all i).

An example

For a Binomial GLM

1. $\{Y_1, ...Y_n\}$ are a set of independent random variables.

2. $\mu_i = E(Y_i) = mp_i$ is related to x_i and β for example through

3. $var(Y_i) =$

(we could also have overdispersion here).

The Quasi-Score function

For one RV Y the **quasi-score** (QS) is defined to be

$$U = U(\mu; Y) = \frac{Y - \mu}{\sigma^2 V(\mu)}$$

.

Properties of U:

$$E(U) = 0;$$

$$var(U) = E(U^2) = \frac{1}{\sigma^2 V(\mu)};$$

$$-E(\frac{\partial U}{\partial \mu}) = \frac{1}{\sigma^2 V(\mu)}.$$

Example: overdispersed Binomial

For the overdispersed Binomial we have

$$V(\mu) = \frac{\mu(m-\mu)}{m}$$

.

Thus the quasi-score is

Defining the Quasi-Likelihood

The **quasi-likelihood** (QL) or more accurately the **log quasi-likelihood** is defined as

$$Q = Q(\mu; Y) = \int_Y^\mu U(t; Y)dt$$
$$= \int_Y^\mu \frac{Y - t}{\sigma^2 V(t)}dt.$$

If this integral exists, it should behave like a log-likelihood function for μ.

Example: overdispersed Binomial

For the overdispersed Binomial the QL is

Quasi-Deviance

The **quasi-deviance** (QD) is defined by

$$\begin{aligned} D(y;\mu) &= 2\sigma^2[Q(y;y) - Q(\mu;y)] \\ &= -2\sigma^2 Q(\mu;y). \end{aligned}$$

Example: for the overdispersed binomial the QD is

The QL and QS for an independent sample

For independent RVs $Y = (Y_1, ...Y_n)^T$ and mean vector $\mu = (\mu_1, .., \mu_n)^T$, the QL for μ based on Y is

$$\begin{aligned} Q(\mu; Y) &= \sum_{i=1}^{n} Q_i(\mu_i, Y_i) \\ &= \sum_{i=1}^{n} \int_{Y_i}^{\mu_i} \frac{Y_i - \mu_i}{\sigma^2 V_i(\mu_i)} dt. \end{aligned}$$

The QS is

$$\begin{aligned} U(\mu; Y) &= \sum_{i=1}^{n} U_i(\mu_i, Y_i) \\ &= \sum_{i=1}^{n} \frac{Y_i - \mu_i}{\sigma^2 V_i(\mu_i)}. \end{aligned}$$

The QD is

$$-2\sum_{i=1}^{n}$$

Example: binomial with overdispersion

Example: Poisson with overdispersion

Calculate the QS, QL and QD for a Poisson model with overdispersion.

Example: Poisson with overdispersion (cont.)

Calculate the QS, QL and QD for a Poisson model with overdispersion.

Estimating β

- Now suppose that $g(\mu_i) = \boldsymbol{x}_i^T \beta$ for each i.
- Remember that the QL for μ_i given Y_i is defined to be

$$Q_i = \int_{Y_i}^{\mu_i} \frac{Y - t}{\sigma^2 V_i(t)} dt.$$

- Taking derivatives we have that the score equation for β_j is

$$\begin{aligned} \frac{\partial Q_i}{\partial \beta_j} &= \frac{\partial Q_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \\ &= U_i \frac{\partial \mu_i}{\partial \beta_j} \\ &= \frac{\partial \mu_i}{\partial \beta_j} \left[\frac{Y_i - \mu_i}{\sigma^2 V_i(\mu_i)} \right]. \end{aligned}$$

- For our set of independent RVs we have that

$$\frac{\partial Q}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_j} \left[\frac{Y_i - \mu_i}{\sigma^2 V_i(\mu_i)} \right]$$

An estimating equation for β

- Note that

$$E\left(\frac{\partial Q_i}{\partial \beta_j}\right) = 0,$$

  for each i and j.

- Matching moments, we obtain the following **estimating equation** for β_j.

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_j}\left[\frac{Y_i - \mu_i}{\sigma^2 V_i(\mu_i)}\right] = 0.$$

- Factoring σ^2 out of this equation, we see that estimation of β_j does not depend on σ^2.

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_j}\left[\frac{Y_i - \mu_i}{V_i(\mu_i)}\right] = 0.$$

- Compare with the ML estimates for exponential families!

Estimating β using IWLS

- Initialize: $\eta = g(\boldsymbol{y})$ (with a suitable adjustment for problems in calculating $g(\boldsymbol{y})$).
- Iteratively calculate the following until changes in β are "small":
  1. $\mu = h(\eta)$ where $h(\cdot)$ is the inverse link function.
  2. Let $\boldsymbol{V}(\mu) = (V_1(\mu_1), ..., V_n(\mu_n))^T$
  3. $\boldsymbol{W} = \text{diag}\left(\left[(g'(\mu))^2 \boldsymbol{V}(\mu)\right]^{-1}\right)$.
  4. $\boldsymbol{z} = \eta + (\boldsymbol{y} - \mu)g'(\mu)$.
  5. $\beta = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z}$.
  6. $\eta = \boldsymbol{X}\beta$.

Asymptotic distribution of β

- Let $\boldsymbol{D}$ be a $n \times p$ matrix with (i,j) element $\partial \mu_i / \partial \beta_j$.
- Let $\boldsymbol{V} = \text{diag}(\boldsymbol{V}(\mu))$.
- Then we wish to solve $\boldsymbol{U}(\hat{\beta}) = 0$,where

$$\boldsymbol{U}(\beta) = \boldsymbol{D}^T \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})/\sigma^2.$$

- Via a Taylor series expansion,

$$0 = \boldsymbol{U}(\hat{\beta}) = \boldsymbol{U}(\beta) + \frac{\partial \boldsymbol{U}(\beta)}{\partial \beta}(\hat{\beta} - \beta).$$

- Thus

$$\sqrt{n}(\hat{\beta} - \beta) = \left[\frac{1}{\sqrt{n}}\boldsymbol{U}(\beta)\right]\left[-\frac{1}{n}\frac{\partial \boldsymbol{U}(\beta)}{\partial \beta}\right]^{-1}.$$

Further properties of the QL derivatives

- We already know that

$$E\left(\frac{\partial Q_i}{\partial \beta_j}\right) = 0.$$

- Now

$$\begin{aligned}
-E\left(\frac{\partial^2 Q_i}{\partial \beta_j \partial \beta_k}\right) &= -E\left(\frac{\partial}{\partial \beta_j}\left\{\frac{\partial \mu_i}{\partial \beta_k}\left[\frac{Y_i - \mu_i}{\sigma^2 V_i(\mu_i)}\right]\right\}\right) \\
&= -E\left((Y_i - \mu_i)\left[\frac{\partial^2 \mu_i}{\partial \beta_j \beta_k} \cdot \frac{1}{\sigma^2 V_i(\mu_i)}\right]\right) \\
&\quad + \frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k}\frac{1}{\sigma^2 V_i(\mu_i)} \\
&= \frac{1}{\sigma^2 V_i(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k},
\end{aligned}$$

since

$$E\left(\frac{\partial}{\partial \beta_j}\left[\frac{Y_i - \mu_i}{\sigma^2 V_i(\mu_i)}\right]\right) =$$

The QL derivatives (cont.)

- We also have that

$$E\left(\frac{\partial Q_i}{\partial \beta_j}\frac{\partial Q_i}{\partial \beta_k}\right) = E(U_i^2)\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k}$$
$$= \frac{1}{\sigma^2 V_i(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k}.$$

- Thus

$$E\left(\frac{\partial Q_i}{\partial \beta_j}\frac{\partial Q_i}{\partial \beta_k}\right) = -E\left(\frac{\partial^2 Q_i}{\partial \beta_j \partial \beta_k}\right)$$

Moments of the terms in the expansion

- Thus $E(\boldsymbol{U}(\beta)) = 0$ and

$$\begin{aligned}\text{cov}(\boldsymbol{U}(\beta)) &= \frac{\boldsymbol{D}^T\boldsymbol{V}^{-1}\text{cov}(\boldsymbol{Y})\boldsymbol{V}^{-T}\boldsymbol{D}}{\sigma^4} \\ &= \frac{\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{V}\sigma^2\boldsymbol{V}^{-T}\boldsymbol{D}}{\sigma^4} \\ &= \frac{\boldsymbol{D}^T V^{-1}\boldsymbol{D}}{\sigma^2}.\end{aligned}$$

- Also by the properties of QL derivatives

$$-E\left(\frac{\partial \boldsymbol{U}(\beta)}{\partial \beta}\right) = \frac{\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}}{\sigma^2}.$$

- With suitable 'hand-waving'

$$\hat{\beta} \approx_d N_p(\beta, \sigma^2(\boldsymbol{D}^T\boldsymbol{V}\boldsymbol{D})^{-1}),$$

and **changes** in the quasi-deviance are asymptotically chi-squared dis-

Estimating σ^2

- As we have done before we estimate σ^2 using

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i-\hat{\mu}_i)^2}{V_i(\hat{\mu}_i)}\\ &= \frac{X^2}{n-p},\end{aligned}$$

  where X^2 is the Pearson X^2 statistic.

- This is also an **estimating equation**:

$$(n-p)\hat{\sigma}^2 - X^2 = 0.$$

The effect of estimating σ^2

- We can replace the asymptotic variance for $\hat{\boldsymbol{\beta}}$ of $\sigma^2(\boldsymbol{D}^T\boldsymbol{V}\boldsymbol{D})^{-1}$ by $\hat{\sigma}^2(\boldsymbol{D}^T\boldsymbol{V}\boldsymbol{D})^{-1}$.
- For the analysis of quasi-deviance table

$$\frac{\text{change in QD}}{\hat{\sigma}^2} \approx_d \chi^2_{p-q},$$

  where $p - q$ denote the number of parameters we added to the model.
- If you are happy about the chi-squared distribution for $\hat{\sigma}^2$ could also use

$$\frac{\text{change in QD}}{(p-q)\hat{\sigma}^2} \approx_d F_{p-q,n-p},$$

Revisiting the solder example

- Let us fit a QL model with a log link and

$$V_i(\mu_i) = \mu_i,$$

  for all i. Assume the variance is $\text{var}(Y_i) = \sigma^2 \mu_i$.

- We use the `quasi` family instead of `Poisson`:

```
model <- glm(skips ~ Opening + Solder + Mask +
PadType + factor(Panel),
data=solder,
family=quasi(link="log", variance="mu"))
```

- For the `quasi` command:
  - Possible **links** are: `logit`, `probit`,`cloglog`, `identity`, `inverse`, `log`, `1/mu^2` and `sqrt`.
  - Possible **variance functions** are: `constant`, `mu(1-mu)`, `mu`, `mu^2` and `mu^3`

The model summary

```
glm(formula = skips ~ Opening + Solder + Mask + PadType + factor(Panel),
    family = quasi(link = "log", variance = "mu"), data = solder)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6615  -1.0868  -0.4407   0.6115   3.9429

Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  -1.21987    0.11623  -10.495   < 2e-16 ***
OpeningM      0.25851    0.08127    3.181  0.001532 **
OpeningS      1.89349    0.06548   28.917   < 2e-16 ***
SolderThin    1.09973    0.04717   23.314   < 2e-16 ***
MaskA3        0.42819    0.09214    4.647  4.02e-06 ***
MaskB3        1.20225    0.08176   14.704   < 2e-16 ***
MaskB6        1.86648    0.07704   24.228   < 2e-16 ***
PadTypeD6    -0.36865    0.08715   -4.230  2.65e-05 ***
PadTypeD7    -0.09844    0.08082   -1.218  0.223640
PadTypeL4     0.26236    0.07412    3.540  0.000427 ***
PadTypeL6    -0.66845    0.09573   -6.982  6.74e-12 ***
PadTypeL7    -0.49021    0.09042   -5.421  8.14e-08 ***
PadTypeL8    -0.27115    0.08472   -3.200  0.001434 **
PadTypeL9    -0.63645    0.09473   -6.718  3.80e-11 ***
```

The model summary (cont.)

```
Coefficients (cont):
                 Estimate Std. Error t value  Pr(>|t|)
PadTypeW4     -0.11000   0.08107   -1.357   0.175252
PadTypeW9     -1.43759   0.12721   -11.301  < 2e-16 ***
factor(Panel)2 0.33352   0.05136    6.494   1.58e-10 ***
factor(Panel)3 0.25440   0.05223    4.871   1.38e-06 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi family taken to be 1.490636)

Null deviance: 6855.7 on 719 degrees of freedom
Residual deviance: 1130.5 on 702 degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

- SEs, t-values and P-values are all adjusted for σ^2.
- The deviances here are quasi-deviances.
- AIC is not defined for a QL model.

Analysis of quasi-deviance

```
Analysis of Deviance Table
Model: quasi, link: log
Response: skips
Terms added sequentially (first to last)

             Df  Deviance Resid. Df Resid. Dev
NULL                           719     6855.7
Opening       2    2524.6       717     4331.1
Solder        1     937.0       716     3394.2
Mask          3    1653.1       713     1741.1
PadType       9     542.5       704     1198.6
factor(Panel) 2      68.1       702     1130.5
```

- Can calculate significance in the usual way.
- OR could use an F-test (see over).
- Diagnostics plots are the usual plots we draw.

Analysis of QD with tests

- If the F test is appropriate use the command:

```
anova(model, test="F")
```

- to get

```
Analysis of Deviance Table
Model: quasi, link: log
Response: skips
Terms added sequentially (first to last)

             Df Deviance Resid. Df Resid. Dev     F      Pr(>F)
NULL                         719      6855.7
Opening       2   2524.6      717      4331.1  846.807  < 2.2e-16 ***
Solder        1    937.0      716      3394.2  628.561  < 2.2e-16 ***
Mask          3   1653.1      713      1741.1  369.662  < 2.2e-16 ***
PadType       9    542.5      704      1198.6  40.435   < 2.2e-16 ***
factor(Panel) 2     68.1      702      1130.5  22.855   2.421e-10 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

An exercise - changing the variance function

- Trying refitting this QL model with the variance function

$$V_i(\mu_i) = \mu_i^2,$$

  for all i, with $\text{var}(Y_i) = \sigma^2\mu_i^2$

- Which model do you prefer?

Empirical variance estimates

- Suppose the assumed variance is incorrect, that is,$var(Y_i) \neq \sigma^2 V_i(\mu_i)$
- We still assume $\{Y_i\}$ are a set of independent RVs.
- In this case we wish to solve $\boldsymbol{U}(\hat{\beta}) = 0$, where

$$\boldsymbol{U}(\beta) = \boldsymbol{D}^T \boldsymbol{V}^{-1}(\boldsymbol{Y} - \mu).$$

(we no longer have σ^2 – think of it being in **V**).

- Similar to previous arguments

$$cov(\hat{\beta}) = \boldsymbol{J}^{-1}\boldsymbol{A}\boldsymbol{J}^{-1}.$$

for $p \times p$ matrices **J** and **A**.

- $\boldsymbol{J}^{-1}\boldsymbol{A}\boldsymbol{J}^{-1}$ is called a **sandwich estimator**.

Components of the empirical variance estimate

- We have that

$$\begin{aligned} \boldsymbol{J} &= E\left(-\frac{\partial U(\beta)}{\partial \beta}\right) \\ &= \boldsymbol{D}^T \boldsymbol{V}^{-1} \boldsymbol{D} \\ &= \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \end{aligned}$$

(from IWLS) and

$$\begin{aligned} \boldsymbol{A} &= cov\left(\boldsymbol{U}(\beta)\right) \\ &= \boldsymbol{D}^T \boldsymbol{V}^{-1} cov(\boldsymbol{Y}) \boldsymbol{V}^{-1} \boldsymbol{D}. \end{aligned}$$

- Since $\{Y_i\}$ independent, we can estimate $cov(\boldsymbol{Y})$ by $(Y_i - \hat{\mu})^2$ along the diagonal elements and zero on the off-diagonals.

Estimating the A matrix

- Hence the (i,i) element of $\boldsymbol{V}^{-1}cov(\boldsymbol{Y})\boldsymbol{V}^{-1}$is
$$\frac{(Y_i-\hat{\mu}_i)^2}{(V_i(\hat{\mu}_i))^2} = \frac{(\gamma_i^p)^2}{V_i(\hat{\mu}_i}$$
where γ_i^pdenotes the ith Pearson residual.
- Also the (i,j) element of **D** is
$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\mu_i}{\eta_j}\frac{\eta_i}{\beta_j} = \frac{1}{g\prime(\hat{\mu}_i)}\boldsymbol{X}_{ij}$$
- Thus the (r,s) element of our estimate of **A** is

$$\begin{aligned}\hat{\boldsymbol{A}}_{rs} &= \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{D}_{\boldsymbol{ir}}\boldsymbol{D}_{\boldsymbol{is}}\frac{(Y_i-\hat{\mu}_i)^2}{(V_i(\hat{\mu}_i))^2} \\ &= \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{X}_{\boldsymbol{ir}}\boldsymbol{X}_{\boldsymbol{is}}\frac{(\gamma_i^p)^2}{(g\prime(\hat{\mu}_i))^2 V_i(\hat{\mu}_i)} \\ &= \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{X}_{\boldsymbol{ir}}\boldsymbol{X}\boldsymbol{is}\omega_i(\gamma_i^p)^2\end{aligned}$$

A function to calculate empirical SEs

```
glm.empirical.SEs <- function (model)
{
## calculate the inverse of the J matrix.
Jinv <- summary(model)$cov.unscaled
## calculate the weights multiplied by the squared residuals
wtr2 <- model$weights * resid(model, type="pearson")^2
## calculate the design matrix
X <- model.matrix(model)
## p = number of parameters in the model
p <- dim(X)[2]
## calculate the estimate of the A matrix
A <- matrix(0, p, p)
for (rr in 1:p)
for (ss in 1:p)
A[rr,ss] <- sum(X[,rr] * X[,ss] * wtr2)
## calculate the sandwich estimate of the variance
## and then take square root of diagonal elements to get SEs
sqrt(diag(Jinv %*% A %*% Jinv))
}
```

Solder dataset: Empirical SEs

- Let‘s reconsider the solder dataset.

```
## read in the data.
solder <- read.table("solder.dat", header=T)
## fit a Poisson model.
model <- glm(skips ~ Opening + Solder + Mask +
PadType + factor(Panel), data=solder,
family=poisson)
## fit a QL model with log link and variance function,
## V(mu) = mu.
model2 <- glm(skips ~ Opening + Solder + Mask +
PadType + factor(Panel), data=solder,
family=quasi(link="log", variance="mu"))
## calculate the empirical SEs
## (it doesn’ t matter if we use model or model2 -- why?)
Emp.SEs <- glm.empirical.SEs(model2)
```

Solder data set: Results

	Estimate	PoissonSE	QuasiSE	EmpirSE
(Intercept)	-1.220	0.095	0.116	0.121
OpeningM	0.259	0.067	0.081	0.090
OpeningS	1.893	0.054	0.065	0.074
SolderThin	1.100	0.039	0.047	0.051
MaskA3	0.428	0.075	0.092	0.091
MaskB3	1.202	0.067	0.082	0.077
MaskB6	1.866	0.063	0.077	0.078
PadTypeD6	-0.369	0.071	0.087	0.083
PadTypeD7	-0.098	0.066	0.081	0.070
PadTypeL4	0.262	0.061	0.074	0.092
PadTypeL6	-0.668	0.078	0.096	0.084
PadTypeL7	-0.490	0.074	0.090	0.105
PadTypeL8	-0.271	0.069	0.085	0.094
PadTypeL9	-0.636	0.078	0.095	0.102
PadTypeW4	-0.110	0.066	0.081	0.082
PadTypeW9	-1.438	0.104	0.127	0.131
factor(Panel)2	0.334	0.042	0.051	0.058
factor(Panel)3	0.254	0.043	0.052	0.059

Solder dataset: Conclusions

Some notes on empirical SEs

- Validity:
  - The empirical SEs are only valid when $n \gg p$. (in our example, n =720 and p =18).
- Testing implications:
  - Empirical SEs can be used to test for the significance of each coefficient, β_j.
  - But, empirical SEs have **no** effect on the (quasi) deviance tests (which are model based).
  - Instead, we can use an empirical score test to test for the significance of more that one β_j.

An empirical score test

- Suppose
  $H_0 : \beta = \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix} versus H_1 : \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$
- Let $U = (U_1(\beta), U_2(\beta))^T$ denote the partition of the score equations according to β_1 and β_2.
- Then the score test statistic is
  $$T = U_2(\hat{\beta^0})^T \left[\hat{cov}(U_2(\hat{\beta^0}))\right]^{-1} U_2(\hat{\beta^0}),$$
  where $\hat{\beta^0} = \left(\hat{\beta_1^0}, 0\right)^T$ is an estimate of β under H_0.
- Under H_0, T has approximately a χ^2_{p-q} distribution, where $p - q$ is the dimension of β_2.