1 Dispersion and deviance residuals

For the Poisson and Binomial models, for a GLM with fitted values $\hat{\mu} = \nabla\Lambda(X\hat{\beta})$ the quantity $D_+(Y, \hat{\mu})$ can be expressed as twice the difference between two maximized log-likelihoods for

$$Y_i \overset{\text{indep}}{\sim} \mathbb{P}_{\eta_i}.$$

The first model is the *saturated* model, i.e. where $\hat{\mu}_i = Y_i$, while the second is the GLM.

Since it is a difference of maximized (nested) log-likelihoods, this difference should have roughly a χ^2 distribution with degrees of freedom $n - \text{rank}(X)$ at least for models such as the count data we modeled using Lindsey's method.

Let's look at the counts data again:

```
%%R
library(sda)
data(singh2002)
labels = singh2002$y
print(summary(labels))
expression_data = singh2002$x
tvals = c()
for (i in 1:6033) {
    tvals = c(tvals, t.test(expression_data[,i] ~ labels, var.equal=TRUE)$statistic)
}
zvals = qnorm(pt(tvals, 100))
```

```
Loading required package: entropy
Loading required package: corpcor
Loading required package: fdrtool
 cancer healthy
     52      50
Warning messages:
1: package 'sda' was built under R version 2.15.3
2: package 'entropy' was built under R version 2.15.3
3: package 'corpcor' was built under R version 2.15.3
```

```
%%R -o eta,counts
bins = c(-Inf, seq(-4,4,length=51), Inf)
counts = c()
for (i in 1:(length(bins)-1)) {
    counts = c(counts, sum((zvals > bins[i]) * (zvals <= bins[i+1])))
}
midpoints = (bins[1:length(bins)-1] + bins[2:length(bins)])/2
counts = counts[2:(length(counts)-1)]
midpoints = midpoints[2:(length(midpoints)-1)]
density.glm = glm(counts ~ poly(midpoints,7), family=poisson(link='log'))
eta = predict(density.glm)
```

```
%%R
print(summary(density.glm))
```

```
Call:
glm(formula = counts ~ poly(midpoints, 7), family = poisson(link = "log"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.79431  -0.69422   0.01274   0.61035   2.08374

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           3.89067    0.03436 113.241  < 2e-16 ***
poly(midpoints, 7)1  -0.19805    0.35327  -0.561    0.575
poly(midpoints, 7)2 -10.89546    0.35131 -31.014  < 2e-16 ***
poly(midpoints, 7)3  -0.14374    0.32855  -0.437    0.662
poly(midpoints, 7)4   1.99022    0.30931   6.434 1.24e-10 ***
poly(midpoints, 7)5   0.01894    0.30917   0.061    0.951
poly(midpoints, 7)6   0.31586    0.20410   1.548    0.122
poly(midpoints, 7)7   0.07490    0.20382   0.367    0.713
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6707.369  on 49  degrees of freedom
Residual deviance:   41.079  on 42  degrees of freedom
AIC: 342.61

Number of Fisher Scoring iterations: 4
```

For this data, the *residual deviance* is the quantity $D_+(Y;\hat{\mu})$

```
eta_saturated = np.log(counts)
dev_resid_sq = 2 * (np.exp(eta) - np.exp(eta_saturated) - counts * (eta-eta_saturated))
dev_resid_sq
```

```
array([ 1.77629347,  0.01742557,  3.45930476,  0.00053596,  0.35270483,
        2.27763749,  0.05016394,  0.1729021 ,  1.49157042,  1.76041477,
        0.00977032,  0.08640754,  0.00470415,  2.06396423,  0.00635758,
        0.60711674,  0.36984542,  0.33563335,  1.93941483,  0.86941851,
        1.2070093 ,  0.155413  ,  0.90682819,  1.38505398,  0.00076262,
        2.75389627,  0.40022683,  0.67820893,  0.59786844,  0.09293117,
        0.07191365,  0.92723779,  0.04781488,  0.08804274,  0.92901589,
        0.37925379,  1.01705818,  0.52260584,  0.55759824,  0.24604725,
        1.34835668,  0.17802013,  0.03493832,  0.00006255,  4.3419532 ,
        0.03166507,  0.00707733,  1.30056171,  3.21955332,  0.00030877])
```

I used the name `dev_resid_sq` above to denote these are squared deviance residuals.

```
dev_resid = np.sign(counts - np.exp(eta)) * np.sqrt(dev_resid_sq)
dev_resid[:10]
```

```
array([-1.3327766 ,  0.13200594,  1.85992063, -0.02315076,  0.59388957,
       -1.50918438, -0.22397308,  0.41581498, -1.22129866,  1.32680623])
```

```
Dplus = dev_resid_sq.sum()
Dplus
```

```
41.078870022329738
```

These deviance residuals are what `R` returns as the residuals of the GLM.

```
%%R
R = resid(density.glm)
print(R[1:10])
print(sum(R^2))
```

```
1           2           3           4           5           6
-1.33277660  0.13200594  1.85992063 -0.02315076  0.59388957 -1.50918438
          7           8           9          10
-0.22397308  0.41581498 -1.22129866  1.32680623
[1] 41.07887
```

Comparing this to a χ^2_{42} distribution, this value is entirely plausible.

```
%%R
print(1-pchisq(41.08,42))
```

```
[1] 0.5112363
```

This phenomenon is different than in OLS regression. In OLS regression

$$D_+(Y;\hat{\mu}) = \frac{1}{\sigma^2}\|Y - X\hat{\beta}\|_2^2 = \frac{1}{\sigma^2}\|P^{\perp}_{\text{col}(X)}Y\|_2^2.$$

The difference is that there is an additional scale parameter to estimate in OLS regression. The binomial and Poisson regression models have no scale parameter.

This is why we see the phrase

```
(Dispersion parameter for poisson family taken to be 1)
```

in `summary(density.glm)`.

The `Null deviance` is

$$D_+(Y;\hat{\mu}_{\text{intercept}})$$

where $\hat{\mu}_{\text{intercept}}$ is the model with only an intercept. That is,

$$\eta_i = \beta_0.$$

1.1 Overdispersion

We can therefore think of the residual deviance as a goodness of fit test. If the model is correct, the residual deviance should be approximately χ^2 with the stated degrees of freedom.

1.1.1 *Exercise: goodness-of-fit test and test of independence*

Men and women in a particular sample were asked whether or not they believe in the afterlife.

	Male	Female
Yes	435	375
No or Undecided	147	134

1. Apply Lindsey's to these $(X_B, X_G) \in \{Y, N\} \times \{M, F\}$ valued random variables and fit a Poisson model to this data under the null hypothesis that X_B is independent of X_G.

2. Compare the residual deviance to the usual Pearson's χ^2 test of independence.

However, in some datasets, the residual deviance is very far off as judged by this scale. Brad's note use this toxoplasmosis incidence data from El Salvador to illustrate this point.

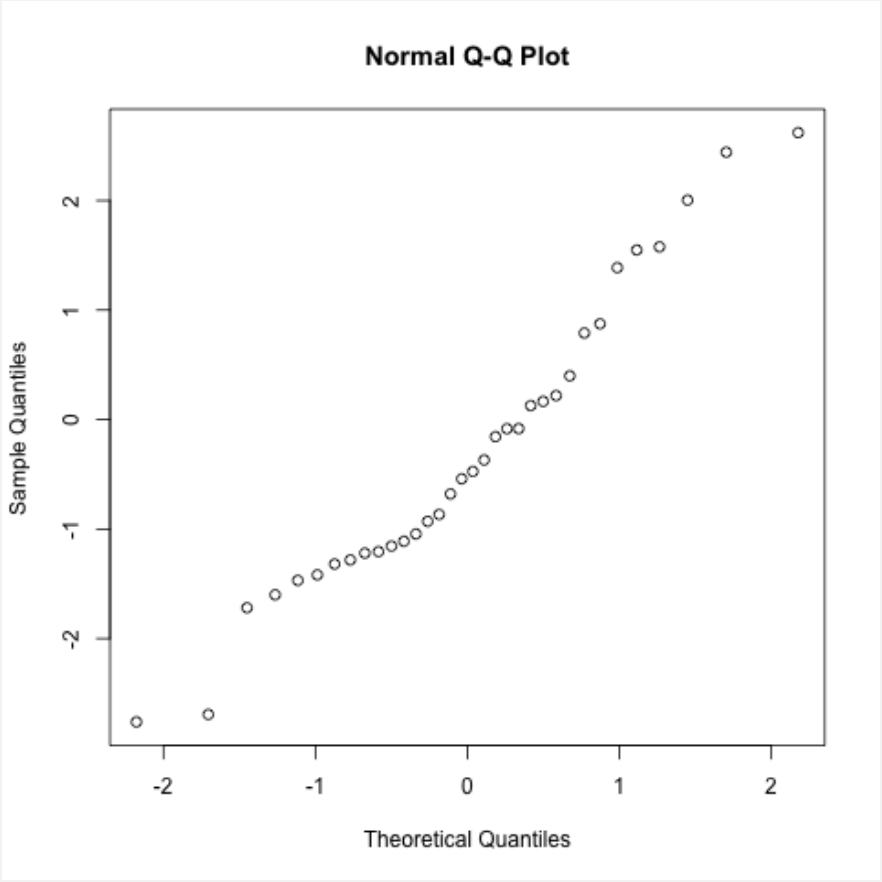
```
%%R
library(SMPracticals)
data(toxo)
toxo$srain = scale(toxo$rain)
toxo.glm = glm(r/m ~ srain + I(srain^2) + I(srain^3), weights=m, data=toxo, family=
    binomial)
print(summary(toxo.glm))
```

```
Call:
glm(formula = r/m ~ srain + I(srain^2) + I(srain^3), family = binomial,
    data = toxo, weights = m)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7620  -1.2166  -0.5079   0.3538   2.6204

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.09939    0.10197   0.975 0.329678
srain       -0.44846    0.15513  -2.891 0.003843 **
I(srain^2)  -0.18727    0.09152  -2.046 0.040743 *
I(srain^3)   0.21342    0.06370   3.351 0.000806 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.212  on 33  degrees of freedom
Residual deviance: 62.635  on 30  degrees of freedom
AIC: 161.33

Number of Fisher Scoring iterations: 3
```

The residual deviance here is 62.63, very large for something nominally χ^2_{30}. There is virtually no chance that a χ^2_{30} would be so large. In this setting, the χ^2_{30} limit would be appropriate if our model were correct and we sampled more and more within each city.

```
%%R
print(1-pnorm(62.63,30))
qqnorm(resid(toxo.glm))
```

```
[1] 0
```



The deviance residuals are generally too large:

```
%R print(sd(resid(toxo.glm)))
```

```
[1] 1.344741
```

So what happened here? What do we do?

It is hard to know exactly why these residuals are too large. It is an indication that our model

$$\mathcal{Y}_\rangle | \mathcal{X} \sim \mathbb{P}_{\eta(X_i)}$$

is incorrect.

Let's refit the model a little differently. The earlier data had been pooled across cities. Let's expand this to a full data set. If our original binomial model was correct, then this new data set will be independent Bernoulli random variables given the covariates. The `rain` part of the model is unnecessary because it is subsumed into the `city` factor.

```
%R -o toxo
rain, nsample, cases, srain = toxo
```

```
bernoulli_cases = []
city = []

city_idx = 0
for n, r, c in zip(nsample, rain, cases):
    bernoulli_cases += c*[1] + (n-c)*[0]
    city += n*[city_idx]
    city_idx += 1
bernoulli_cases = np.array(bernoulli_cases)
%R -i bernoulli_cases,city
```

```
%%R
bernoulli_cases = as.numeric(bernoulli_cases)
expanded_rain = as.numeric(expanded_rain)
city = as.factor(city)
expanded.glm = glm(bernoulli_cases ~ poly(expanded_rain,3), family=binomial())
print(summary(expanded.glm))
```

```
Call:
glm(formula = bernoulli_cases ~ poly(expanded_rain, 3), family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3750  -1.2085   0.9919   1.1007   1.4433

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.04289    0.07641   0.561 0.574628
poly(expanded_rain, 3)1 -0.67501    2.03019  -0.332 0.739523
poly(expanded_rain, 3)2 -0.01739    2.03838  -0.009 0.993191
poly(expanded_rain, 3)3  6.82977    2.03841   3.351 0.000807 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 965.92  on 696  degrees of freedom
Residual deviance: 954.35  on 693  degrees of freedom
AIC: 962.35

Number of Fisher Scoring iterations: 4
```

```
%%R
print(anova(expanded.glm))
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: bernoulli_cases

Terms added sequentially (first to last)


                      Df Deviance Resid. Df Resid. Dev
NULL                                    696     965.92
poly(expanded_rain, 3)  3   11.577       693     954.35
city                   30   62.635       663     891.71
```

1.1.2 *Exercise*

Why does the deviance decrease for adding `city` as a factor equal the residual deviance from the original model?

One reason for the model to be off might be that our assumption of conditional independence given our covariates might be incorrect. For instance, we might not truly have independent samples within each city, so there may be some unknown clustering in our data. That is, the binomial model for the counts we observed in the model `toxo.glm` may not be appropriate.

1.2 Quasilikelihood

The quasilikelihood approach assumes that the mean model is correct. That is, if $\mu_i = \mathbb{E}(\mathcal{Y}_i|\mathcal{X})$, then we continue to assume

$$g(\mu_i) = x_i^T \beta.$$

If the GLM is correct, then the variance is tied to the mean by

$$\begin{aligned}\text{Var}(\mathcal{Y}_i|\mathcal{X}) &= \text{Var}(\mathcal{Y}_i|\mathcal{X}_i) \\ &= \text{Var}_{F(X_i^T\beta)}(Y) \\ &= \ddot{\Lambda}(F(x_i^T\beta)).\end{aligned}$$

The quasilikelihood assumption is that

$$\text{Var}(\mathcal{Y}_i|\mathcal{X}) = \phi^{-1} \cdot \ddot{\Lambda}(F(x_i^T\beta)) = \phi^{-1} \cdot \text{Var}_{\eta(\mu_i)}(Y)$$

The model is fit using the *quasi-likelihood* for η which we can think of, formally, as

$$\ell_{\text{quasi}}(\beta) = \phi^{-1} \cdot \ell(\beta)$$

where $\ell(\beta)$ is the likelihood assuming the original GLM was correct.

Strictly speaking, the quasi-likelihood is a function whose score is the gradient of what I called $\nabla\ell_{\text{quasi}}(\beta)$:

$$\nabla\ell_{\text{quasi}}(\beta) = \phi^{-1} \cdot X^T \nabla F(X\beta) \left[\nabla\Lambda^{(n)}(F(X\beta)) - Y\right]$$

and the formal calculation goes on to the (quasi) Fisher information

$$\text{“}\mathbb{E}_\beta(\nabla \ell_{\text{quasi}}(\beta)\nabla \ell_{\text{quasi}}(\beta)^T)'' = \phi^{-1} \cdot X^T \nabla^2 \Lambda^{(n)}(F(X\beta))X.$$

Leading us to conclude

$$\hat{\beta} \approx N\left(\beta, \phi \cdot \left(X^T \nabla^2 \Lambda^{(n)}(F(X\beta))X\right)^{-1}\right).$$

I say formally here, because the quasilikelihood is not a true likelihood. For one thing, it would imply that the means and variances of Y would multiplied by ϕ while we know that Y is Bernoulli in the expanded dataset.

1.2.1 *Exercise: quasi-score*

Suppose that, marginally, $Y_i \sim \text{Bernoulli}(\pi(X_i)), 1 \leq i \leq n$ but the joint distribution is unkown and

$$\pi(x_i) = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}.$$

Show that

$$\mathbb{E}(\nabla \ell_{\text{quasi}}(\beta)) = 0.$$

A common estimate of ϕ is

$$\hat{\phi} = \frac{1}{n-p} D_+(Y; \hat{\mu}).$$

Another is Pearson's X^2

$$X^2 = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\text{Var}_{\eta(\hat{\mu}_i)}(Y_i)}$$

```
%%R
phihat = sum(resid(toxo.glm)^2) / 30
print(summary(toxo.glm, dispersion=phihat))
```

```
Call:
glm(formula = r/m ~ srain + I(srain^2) + I(srain^3), family = binomial,
    data = toxo, weights = m)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7620  -1.2166  -0.5079   0.3538   2.6204

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.09939    0.14733   0.675   0.4999
srain       -0.44846    0.22416  -2.001   0.0454 *
I(srain^2)  -0.18727    0.13224  -1.416   0.1568
I(srain^3)   0.21342    0.09204   2.319   0.0204 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 2.08782)
```

```
    Null deviance: 74.212  on 33  degrees of freedom
Residual deviance: 62.635  on 30  degrees of freedom
AIC: 161.33

Number of Fisher Scoring iterations: 3
```

R has a special family for the *quasilikelihood*

```
%%R
quasitoxo.glm = glm(r/m ~ srain + I(srain^2) + I(srain^3) , weights=m, family=
    quasibinomial(link='logit'), data=toxo)
print(summary(quasitoxo.glm))
```

```
Call:
glm(formula = r/m ~ srain + I(srain^2) + I(srain^3), family = quasibinomial(link = "logit"),
    data = toxo, weights = m)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7620  -1.2166  -0.5079   0.3538   2.6204

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09939    0.14204   0.700   0.4895
srain       -0.44846    0.21610  -2.075   0.0466 *
I(srain^2)  -0.18727    0.12749  -1.469   0.1523
I(srain^3)   0.21342    0.08873   2.405   0.0225 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.940446)

    Null deviance: 74.212  on 33  degrees of freedom
Residual deviance: 62.635  on 30  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 3
```

I tried, but was unsuccessful at computing their scale parameter but I believe it is Pearson's X^2 with weights incorporated into the formula.

```
%%R
Y = toxo$r / toxo$m
print(sum(toxo$m*(Y-mu)^2 / ((mu*(1-mu)))))
print(sum(toxo$m))
```

```
[1] 248.6024
[1] 697
```

1.3 General form of a GLM

Having introduced quasilikelihood, which is determined by a link function and a *quasi-family* we see the main ingredients of a *GLM* for a conditional distribution

$$\mathcal{Y} \in \mathbb{R}^n | \mathcal{X} \in \mathbb{R}^{n \times p}$$

1. A link function: suppose we know that $\mathcal{Y}_i \in I$ for some interval I. Then, any differentiable invertible function $g : I \to \mathbb{R}$ can be used as a link function

$$g(\mathbb{E}(\mathcal{Y}_i|\mathcal{X})) = x_i^T \beta.$$

2. A variance function, $V : I \to [0, +\infty)$.

Our original GLMs, the Poisson and binomial came from exponential families and we used the canonical link. As its name implies, the canonical link provided our link function with $I = \mathcal{M}$. The variance function also comes from the exponential family

$$V_\mu = \ddot{\Lambda}(\eta(\mu)) = \mathrm{Var}_{\eta(\mu)}(Y).$$

1.4 Double exponential families

The quasilikelihood approach has the downside that using it as our objective function does not correspond to any true likelihood, not even a conditional likelihood as was the case for pseudolikelihood.

The double exponential family approach is a genuine exponential family approach to modelling overdispersion.

Given an exponential family, $\mathbb{P}_\eta$ with sufficient statistic y and carrier measure m, the double exponential family is an exponential family with the *same* carrier measure

$$\frac{d\mathbb{Q}_{\zeta,\theta}}{dm} = C(\zeta, \theta) \cdot \exp\left(\zeta^T y + \theta(\Lambda(\eta(y)) - \eta(y)^T y)\right), \qquad \theta \geq 0$$

with

$$\eta(y) = \nabla \Lambda^*(y).$$

Above,

$$\log C(\zeta, \theta) = -\log\left[\int_\Omega \exp\left(\zeta^T y + \theta(\Lambda(\eta(y)) - \eta(y)^T y)\right)\ m(dy)\right]$$

is the CGF of the family $\mathbb{Q}_{\zeta,\theta}$.

1.4.1 *Exercise: double exponential families*

What is the domain of the exponential family $\mathbb{Q}_{\zeta,\theta}$?

The only change from the original family is the addition of a new sufficient statistic $\Lambda(\eta(y)) - \eta(y)^T y$. Of course this changes the CGF, which is reflected in $C(\zeta, \theta)$ above.

We might also consider adding an index n corresponding to repeated sampling. In the toxoplasmosis example, this repeated sampling will be the number of samples within each city.

The density therefore has the form

$$\begin{aligned}\frac{d\mathbb{Q}_{\zeta,\theta,n}}{dm} &= C(\zeta,\theta,n)\cdot\theta^{1/2}e^{n\left[\eta(\bar{y})^T\bar{y}-\Lambda(\eta(\bar{y}))-\theta\Lambda(\zeta/\theta)\right]} \\ &\qquad \exp\left[n\left(\zeta^T\bar{y}+\theta\cdot(\Lambda(\eta(\bar{y}))-\eta(\bar{y})^T\bar{y})\right)\right] \\ &= C(\zeta,\theta,n)\;\theta^{1/2}e^{n(1-\theta)\left(\eta(\bar{y})^T\bar{y}-\Lambda(\eta(\bar{y}))\right)}e^{n\theta\left((\zeta/\theta)^T y-\Lambda(\zeta/\theta)\right)} \\ &= C(\zeta,\theta,n)\;\theta^{1/2}\cdot e^{-n\theta\cdot D(\eta(\bar{y});\zeta/\theta)/2}\cdot e^{n\left[\eta(\bar{y})^T\bar{y}-\Lambda(\bar{y})\right]}\end{aligned}$$

where the reference measure is replaced with the reference measure of

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

It is fair to ask why we have introduced the term $\theta^{1/2}e^{n\left[\eta(\bar{y})^T\bar{y}-\Lambda(\bar{y})-\theta\Lambda(\zeta/\theta)\right]}$ above. This both changes the reference measure as well as the $C(\zeta,\theta)$ above.

There are two justifications. Firstly, under the new reference measure, the density is proportional to

$$\theta^{1/2}e^{-n\theta\cdot D(\eta(\bar{y});\eta)/2}=\theta^{1/2}e^{-n\theta\cdot\tilde{D}(\bar{y};\mu(\eta))/2}.$$

The second is due to the fact that, with this choice

$$C(\zeta,\theta,n)\overset{n\to\infty}{\approx}1.$$

See Brad's above paper for more details. This means that

$$f_{\zeta,\theta,n}(\bar{y})=\theta^{1/2}\cdot e^{-n\theta\cdot D(\eta(\bar{y});\zeta/\theta)/2}\cdot e^{n\left[\eta(\bar{y})^T\bar{y}-\Lambda(\bar{y})\right]}$$

is almost a probability density with respect to the reference measure of $\bar{y}$. Let's call it a quasi-density. If it was a density, it would be the density of the exponential family $\mathbb{Q}_{\zeta,\theta,n}$.

1.4.2 *Exercise: double exponential quasidensities*

1. Plot the quasidensity $f_{\zeta,\theta,n}(\bar{y})$ for the Poisson family with $n=10,\zeta=1$ for various values of θ.
2. Vary n over some range. Is the justification $C(\zeta,\theta,n)\approx 1$ reasonable?

1.4.3 Mean and variance

If $f_{\zeta,\theta,n}$ was a density, it would be the density of the exponential family $\mathbb{Q}_{\zeta,\theta,n}$.

Assuming it is a density yields the means and variances of the sufficient statistics in this exponential family

$$\begin{aligned}\mathbb{E}^{(n)}_{\zeta,\theta}\left(n\bar{y}\right) &\approx n\nabla\Lambda(\zeta/\theta) \\ \mathbb{E}^{(n)}_{\zeta,\theta}\left(n\left[\Lambda(\eta(\bar{y}))-\eta(\bar{y})^T\bar{y}\right]\right) &\approx n\left[\Lambda(\zeta/\theta)-\nabla\Lambda(\zeta/\theta)^T\zeta/\theta\right]-\frac{1}{2\theta}.\end{aligned}$$

As for variance,

$$\begin{aligned}\mathrm{Var}^{(n)}_{\zeta,\theta}\left(n\bar{y}\right) &\approx \frac{n}{\theta}\nabla^2\Lambda(\zeta/\theta) \\ \mathrm{Var}^{(n)}_{\zeta,\theta}\left(n\left[\Lambda(\eta(\bar{y}))-\eta(\bar{y})^T\bar{y}\right]\right) &\approx \frac{n}{\theta}\nabla\Lambda(\zeta/\theta)(\zeta/\theta,\zeta/\theta)+\frac{1}{2\theta^2} \\ \mathrm{Cov}^{(n)}_{\zeta,\theta}\left(n\bar{y},n\left[\Lambda(\eta(\bar{y}))-\eta(\bar{y})^T\bar{y}\right]\right) &\approx -n\nabla^2\Lambda(\zeta/\theta)\zeta/\theta^2.\end{aligned}$$

1.4.4 Interpretation of the parameters

Inspection of the density above shows that, for $\theta < 1$ the density is less concentrated than at $\theta = 1$. Similarly, for $\theta > 1$, the density is more concentrated than at $\theta = 1$.

Hence, θ is inversely related to dispersion: $\theta < 1$ indicates overdispersion, while $\theta > 1$ indicates underdispersion.

The parameter ζ/θ appears as an argument to Λ, hence we can think of ζ/θ as η in the model $\theta = 1$. If we parameterize this family by (η, θ) then it is no longer an exponential family in the sense we've been using the term because the parameters do not appear linearly. It is a *curved exponential family*.

Nevertheless, this (η, θ) parametrization is essentially the one that Brad uses in his paper (actually he uses $\mu = \nabla\Lambda(\zeta/\theta)$).

In the $\zeta = \eta\theta$ parameterization, we see

$$\begin{aligned}\frac{d\mathbb{Q}_{\eta\theta,\theta,n}}{dm} &= C(\theta\eta, \theta, n) \cdot \theta^{1/2} e^{-n\theta \cdot D(\eta(\bar{y});\eta)/2} \cdot e^{n\left[\eta(\bar{y})^T \bar{y} - \Lambda(\bar{y})\right]} \\ &= C(\theta\eta, \theta, n) \cdot \theta^{1/2} e^{(1-\theta)n\left(\eta(\bar{y})^T \bar{y} - \Lambda(\eta(\bar{y}))\right)} e^{\theta n\left(\eta^T \bar{y} - \Lambda(\eta)\right)}.\end{aligned}$$

1.4.5 *Exercise: Fisher information of double exponential familes*

1. Compute the Fisher information of $(\eta, \theta) = (\zeta/\theta, \theta)$ in the double exponential family. (You can ignore the normalization constant).

2. Argue that this implies that $(\hat{\eta}, \hat{\theta})$ are asymptotically independent as $n \to \infty$.

1.4.6 Estimation in double exponential families

As $C(\zeta, \theta, n)$ is very close to 1 for n sufficiently large, one may effectively ignore C in MLE computations.

The score vector is

$$\nabla \log \frac{d\mathbb{Q}_{\zeta,\theta,n}}{dm} \approx \begin{pmatrix} -\frac{n}{2}\nabla_\eta D(\eta(\bar{y});\eta)\Big|_{\eta=\zeta/\theta} \\ \frac{1}{2\theta} - \frac{1}{2}D(\eta(\bar{y});\zeta/\theta) + \left(\frac{1}{2}\nabla_\eta D(\eta(\bar{y});\eta)\Big|_{\eta=\zeta/\theta}\right)^T \zeta/\theta \end{pmatrix}$$

Above, the approximation ignore $\nabla C(\zeta, \theta, n)$ which we take to be approximately 0.

We see, then that for the MLE $(\hat{\zeta}, \hat{\theta})$

$$\begin{aligned}\hat{\zeta}/\hat{\theta} &= \nabla\Lambda^*(\bar{y}) \\ \frac{1}{\hat{\theta}} &= n \cdot D(\eta(\bar{y}); \hat{\zeta}/\hat{\theta}) \\ &= D^{(n)}(y, \hat{\zeta}/\hat{\theta}).\end{aligned}$$

1.5 Regression and double exponential families

Suppose now we form a regression model using double exponential families with repeated measurements. In this model, we assume that, as in the toxoplasmosis case, each outcome can be thought of as repeated measurements for some fixed value of the covariates.

We follow the setup of our generalized linear model. That is, we assume that, given $\mathcal{X}_i, 1 \leq i \leq N$ we observe

$$\mathcal{Y}_i | \mathcal{X} \overset{\text{indep}}{\sim} \mathbb{Q}_{\zeta_i, \theta_i, n_i} = \mathbb{Q}_{\eta_i \theta_i, \theta_i, n_i}$$

As the θ_i's are some version of a dispersion parameter, we of course have the choice to assume that they are constant for all i, leading to the model

$$\mathcal{Y}_i | \mathcal{X} \overset{\text{indep}}{\sim} \mathbb{Q}_{\zeta_i, \theta, n_i} = \mathbb{Q}_{\eta_i \theta, \theta, n_i}.$$

The natural linear model to consider is

$$\eta_i = x_i^T \beta = \zeta_i / \theta.$$

This leads to a likelihood

$$\ell(\beta, \theta) = \frac{N}{2} \log \theta + \frac{\theta}{2} \sum_{i=1}^{N} n_i D(\eta(\bar{y}); x_i^T \beta).$$

1.5.1 *Exercise: score for double exponential linear models*

1. Compute the score equations of the above linear model.
2. Show that $\hat{\beta}$ is exactly the same as the quasilikelihood estimator of β.
3. Show that $\hat{\theta}$ is *almost* the same as the quasilikelihood estimator of θ. What's different?
4. Compute the Fisher information of (β, θ) in this linear model. What advantage does this model have over the quasilikelihood model? (Hint: can you compute the Fisher information of the pair (β, θ) in the quasilikelihood setup?)

1.5.2 *Exercise: double exponential linear models*

1. Show the linear model we've formed for repeated measurements is not an exponential family (in the sense we've been using the term exponential family)?
2. Modify the model so that it is an exponential family.

1.5.3 *Exercise: modelling dispersion*

In the linear model we introduced above we made the assumption that $\theta_i = \theta$, but we might want to model θ as a function of x as well. Suppose that $\eta_i = x_i^T \alpha$.

1. Form a regression model for $\mathcal{Y}_i | \mathcal{X}$ that allows both (ζ, θ) to depend on some covariates. Choose your model so that it is a genuine exponential family.
2. Write out the score and Fisher information for this model.

3. Fit this model to the toxoplasmosis data where the effect on ζ is assumed to be cubic while the effect on θ is assumed to be linear (plus a constant). (Don't forget the constraint $\theta \geq 0$.)

4. This model is some sort of random effects model. Suppose that the original family we started with was the normal means model (i.e. reference measure $e^{-x^2/2}$ on $\mathbb{R}$ and sufficient statistic x). Describe this model of the variance for the random means model.

1.5.4 Reduced sample size

The quasidensity calculations above show that, under $\mathbb{Q}_{\zeta,\theta,n}$ we have

$$\begin{aligned}
\mathbb{E}^{(n)}_{\zeta,\theta}(\bar{y}) &\approx \nabla\Lambda(\zeta/\theta) \\
&= \nabla\Lambda(\eta) \\
\mathrm{Var}^{(n)}_{\zeta,\theta}(\bar{y}) &\approx \frac{1}{n\theta}\nabla^2\Lambda(\zeta/\theta) \\
&= \frac{1}{n\theta}\nabla^2\Lambda(\eta).
\end{aligned}$$

So, it seems as if the effect of introducing θ effectively reduces the sample size by a factor of θ.

The double exponential family approach with a common θ is almost like changing n_i to $n_i\theta$, but it has the advantage that it does so in an exponential family way.

Because the reference measure is the *same* as in the original family, one isn't changing the parameter space with θ. That is, for Binomial data, changing n_i to $n_i\theta$ changes the support of $\bar{y}_i$ because it changes the denominator.

1.5.5 *Exercise: Binomial double exponential family quasidensity*

Suppose the original family we start with is Bernoulli(π).

1. What is the quasidensity $f_{\zeta,\theta,n}$ for $\bar{y}$? (You can ignore the normalizing constant.)

2. Make a plot of the quasidensity for $n = 16, \pi = 0.4, \theta = 0.5$.

3. Compare this quasidensity to the density of $\bar{y}$ if we assume the reduced sample size $\tilde{n} = n\theta = 8$.