Correction **Assignment**

Exercise 1.

a) **[Seen, 4 points]** Taylor expansion of the first derivative of the score function, gives

$$\frac{\partial\eta^{\mathrm{T}}}{\partial\beta}u(\beta)+\left\{\sum_{j=1}^{n}\frac{\partial\eta^{\mathrm{T}}}{\partial\beta}\frac{\partial^2\ell_j}{\partial\eta_j^2}\frac{\partial\eta^{\mathrm{T}}}{\partial\beta}+\sum_{j=1}^{n}\frac{\partial^2\eta_j}{\partial\beta\partial\beta^{\mathrm{T}}}u_j(\beta)\right\}(\widehat{\beta}-\beta),$$

where $u(\beta)=\partial\ell/\partial\eta$ and ℓ is a log-likelihood function with parameter η. Suppose that $\ell_j(\eta_j)=\log f(y_j;\eta_j)$, where the density f is regular for maximum likelihood estimation.

It is convenient to replace the quantity in brace by its expectation, which leads to

$$\frac{\partial\eta^T}{\partial\beta}u(\beta)+\left[\sum_{j=1}^{n}\frac{\partial\eta^T}{\partial\beta}\mathrm{E}\left(\frac{\partial^2\ell_j}{\partial\eta_j^2}\right)\frac{\partial\eta^T}{\partial\beta}+\sum_{j=1}^{n}\frac{\partial^2\eta_j}{\partial\beta\partial\beta^T}\mathrm{E}\left\{u_j(\beta)\right\}\right](\widehat{\beta}-\beta)=0.$$

Then the equality simplifies to

$$X^Tu(\beta)+\left[\sum_{j=1}^{n}X^T\mathrm{E}\left(\frac{\partial^2\ell_j}{\partial\eta_j^2}\right)X\right](\widehat{\beta}-\beta)=0.$$

with $X=\frac{\partial\eta^T}{\partial\beta}$, and which leads to

$$\widehat{\beta}=\beta+\left\{X^TWX\right\}^{-1}X^Tu(\beta)=\left\{X^TWX\right\}^{-1}X^TWz,$$

where $z=X\beta+W^{-1}u(\beta)$ and W is a diagonal matrix with terms $\mathrm{E}\left(-\frac{\partial^2\ell_j}{\partial\eta_j^2}\right)$. Then we use this formula for setpwise optimization starting from a good β.

b) **[Seen, 2 points]** We use profile log-likelihood for β_r, with fixed β_{-r} parameters. We know from the likelihood ratio statistics that

$$W(\beta_1)=2\left\{\ell(\widehat{\beta}_1)-\ell(\beta_1)\right\}\sim\chi_1^2.$$

Thus the set of plausible values with a $(1-2\alpha)$ level is

$$\left\{\beta_1:\ell(\beta_r)>\ell(\widehat{\beta})-\tfrac{1}{2}c_1(1-2\alpha)\right\},$$

where $c_1(1-2\alpha)$ is the $(1-2\alpha)$ quantile of the χ_1^2 distribution. We can then compute the corresponding confidence intervals, if we specify a value for α. In this case, testing can be done using the deviance statistics of nested models.

Also possible:

$$\widehat{\beta}_r+2V_{rr}^{1/2},\quad\frac{\widehat{\beta}_r-\beta}{V_{rr}^{1/2}}\sim N(0,1).$$

c) **[Unseen, 4 points]** We suppose that

$$V \sim N\left(\tfrac{Uc}{K+c}, \sigma^2\right),$$

thus we have

$$\tfrac{\partial \eta}{\partial U} = \tfrac{c}{K+c}, \quad \tfrac{\partial \eta}{\partial K} = \tfrac{-cU}{(K+c)^2},$$

and

$$\ell_i(y_i; \eta_i) = -\log(\sqrt{2\pi\sigma^2}) - \tfrac{1}{2\sigma^2}(y_i - \eta_i)^2.$$

This leads to $W_{ii} = 1/\sigma^2$, $i = 1, \ldots, n$, which means that W is proportional to the identity matrix. We can deduce that

$$z = X^T \begin{bmatrix} U \\ K \end{bmatrix} + \{Y - \eta(K, U)\},$$

with,

$$\widehat{\beta} = \left\{X^T W X\right\}^{-1} X^T W z,$$

where X is a $2 \times n$ matrix with columns $\left[\frac{c_j}{K+c_j} \; \frac{-c_j U}{(K+c_j)^2}\right]^T$.

Bonus: mention the case where $U = 0$ then the observation matrix is singular ...

Exercise 2.

a) **[Seen, 3 points]** The mean can be seen as a smooth function of the covariates, more precisely a polynomial function of degree p as well as a non parametric smooth component:

- y is the vector of observation,
- γ is the vectors of parameter $(\beta_0, \beta_1, \ldots, \beta_b, b_1, \cdot, b_k)$,
- B is the matrix of covariates with $\{1, x, \ldots, x^p, (x - \kappa_1)^p, \ldots, (x - \kappa_k)^p\}$,
- α is a smoothing parameter,
- D is a diagonal matrix with p zeros and k times 1.

b) **[Seen, 2 points]** We take partial derivative with regard to γ in the previous equation and we get

$$2B^T(y - B\gamma) + 2\alpha D\gamma = 0.$$

This gives

$$\gamma_\alpha = (B^T B + \alpha D)^{-1} B^T y$$

and thus,

$$\widehat{y} = B(B^T B + \alpha D)^{-1} B^T y = S_\alpha y.$$

c) **[Seen, 2 points]** Let M and N be $q \times q$ matrices, and suppose that $(N + \alpha M)^{-1}$ exists for some $\alpha > 0$. Let η be an eigenvalue of $(N + \alpha M)^{-1} N$. Then if N is invertible, then

$$\begin{aligned}(N + \alpha M)^{-1} A &= (N^{1/2} N^{1/2} + \alpha M)^{-1} A \\ &= N^{-1/2}(I + \alpha N^{-1/2} M N^{-1/2})^{-1} N^{-1/2} N \\ &= N^{-1/2}(I + \alpha N^{-1/2} M N^{-1/2})^{-1} N^{1/2},\end{aligned}$$

which gives

$$\eta = \frac{1}{1 + \alpha\eta''},$$

where η'' is an eigenvalue of $N^{-1/2} M N^{-1/2}$.

d) **[Seen, 3 points]** With the result of the previous question, we have

$$tr(S_\alpha) = \sum_{j=1}^{n} \tfrac{1}{1+\alpha\eta_j},$$

where η_j are the eigenvalues of $(B^TB)^{-1/2}D(B^TB)^{-1/2}$. We see that $\eta_1 = \cdots = \eta_{p+1} = 0$ and we suppose $0 < \eta_{p+2} \leqslant \cdots \leqslant \eta_{p+1+k}$. Thus we get

$$tr(S_\alpha) = p + 1 + \sum_{j=p+2}^{n} \tfrac{1}{1+\alpha\eta_j},$$

and thus

$$p + 1 \leqslant tr(S_\alpha) \leqslant p + 1 + K.$$

Monotony is straight forward. $tr(S_\alpha)$ can be seen as the equivalent degree of freedom. When $\alpha = 0$, then $tr(S_\alpha) = p + 1 + K$, which correponds to the case with no smoothing, i.e. classical polynomial regression with function basis. When $\alpha = \infty$, then $tr(S_\alpha) = p + 1$ and we have a classical polynomial regression.

Exercise 3.

a) **[Seen, 3 points]** Suppose that Y has a continuous density; if not the argument below is the same, except that integral signs are replaced by summations.

Let $\Omega_\theta = \{\theta : b(\theta) < \infty\}$.

We have

$$M_Y(t) = \mathrm{E}\{\exp(tY)\} = \int e^{ty} \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\} \mathrm{d}y = \int \exp\left\{\frac{y(\theta + t\phi) - b(\theta)}{\phi} + c(y;\phi)\right\} \mathrm{d}y.$$

If $\theta + t\phi \in \Omega_\theta$, then

$$\int \exp\left\{\frac{y(\theta + t\phi) - b(\theta + t\phi)}{\phi} + c(y;\phi)\right\} \mathrm{d}y = 1,$$

so

$$M_Y(t) = \mathrm{E}\{\exp(tY)\} = \exp\left[\{b(\theta + t\phi) - b(\theta)\}/\phi\right].$$

Hence the cumulant-generating function of Y is

$$K_Y(t) = \log M_Y(t) = \{b(\theta + t\phi) - b(\theta)\}/\phi,$$

and differentiating twice with respect to t and setting $t = 0$ yields

$$K_Y'(t)\big|_{t=0} = b'(\theta), \quad K_Y''(t)\big|_{t=0} = \phi b''(\theta).$$

Since $b(\theta)$ is strictly convex on Ω_θ, $b'(\theta)$ is a monotonic increasing function of θ, so $b'^{-1}(\cdot)$ exists and is itself monotonic, so $V(\mu) = b''\{b'^{-1}(\mu)\}$ is well-defined.

b) **[Seen, 2 points]** The generalized linear model extends classical linear normal model to

- Y has density/mass function

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\}, \quad y \in \mathcal{Y}, \theta \in \Omega_\theta, \phi > 0, \tag{1}$$

  where
  - $\mathcal{Y}$ is the support of Y,
  - Ω_θ is the parameter space of valid values for $\theta \equiv \theta(\eta)$, and
  - the $\beta dispersion parameter$ ϕ is often known;

- with the link function: $\eta = g(\mu) = \theta = b'^{-1}(\mu)$, where μ is the mean. The link function is monotonic, smooth and links $X^T\beta = g\{\mathrm{E}(Y)\}$.
- and variance function: $\mathrm{var}(Y) = \phi V(\mu)$,
- Only V and g appears in the algorithm.

c) **[Unseen, 5 points]** We have

$$\begin{aligned}\Pr(Z = z) &= \{(1-\gamma)\pi + \gamma(1-\pi)\}^z\{1-(1-\gamma)\pi - \gamma(1-\pi)\}^{1-z}\\ &= \exp\left[z\log\left\{\tfrac{\pi(1-2\gamma)+\gamma}{1-\pi(1-2\gamma)-\gamma}\right\} + \log\left\{1-\pi(1-2\gamma)-\gamma\right\}\right].\end{aligned}$$

Thus we have a glm with $\theta = \log\left\{\frac{\pi(1-2\gamma)+\gamma}{1-\pi(1-2\gamma)-\gamma}\right\}$. The corresponding link function is:

$$\begin{aligned}\mathrm{E}(Z_j) &= \mathrm{E}\{(1-I_j)Y_k\} + \mathrm{E}\{I_j(1-Y_j)\}\\ &= (1-\gamma)\pi + \gamma(1-\pi)\\ &= \pi(1-2\gamma)+\gamma,\\ \mu &= \tfrac{e^\eta}{1+e^\eta}(1-2\gamma)+\gamma.\end{aligned}$$

Also

$$\pi = \tfrac{e^\eta}{1+e^\eta} = \tfrac{\mu-\gamma}{(1-2\gamma)}.$$

This leads to

$$\begin{aligned}\eta &= \log\left(\frac{\frac{\mu-\gamma}{(1-2\gamma)}}{1-\frac{\mu-\gamma}{(1-2\gamma)}}\right)\\ &= \log\left(\tfrac{\mu-\gamma}{1-2\gamma-\mu+\gamma}\right)\\ &= \log\left(\tfrac{\mu-\gamma}{1-\gamma-\mu}\right).\end{aligned}$$

and we get the link function. The $b(\theta)$ function is

$$b(\theta) = \log\left\{1-\pi(1-2\gamma)-\gamma\right\} = \log\left\{1-\mu\right\}.$$

This gives,

$$b''(\theta) = \mu(1-\mu)$$

and thus

$$V(\mu) = \left\{\tfrac{e^\eta}{1+e^\eta}(1-2\gamma)+\gamma\right\}\left\{1-\tfrac{e^\eta}{1+e^\eta}(1-2\gamma)+\gamma\right\}$$

If $\gamma = 0.5$, $\mu = \gamma$ and then we cannot estimate π. To estimate γ, we can use a profile log-likelihood with a grid search.

Exercise 4.

a) **[Seen, 3 points]** We can formulate the mixed model as

$$y \mid b \sim \mathcal{N}_n(X\beta + Zb, \sigma^2 I_n), \quad b \sim \mathcal{N}_q\{0, \sigma^2 Q(\psi)\},$$

and then Y is normally distributed and

$$\mathrm{E}(Y) = \mathrm{E}_b\mathrm{E}(y|b) = X\beta + \mathrm{E}_b(Zb) = X\beta,$$

also,

$$\mathrm{var}(Y) = \mathrm{E}_b\mathrm{var}(y|b) + \mathrm{var}\mathrm{E}_b(y|b) = \sigma^2\{I_n + ZQ(\psi)Z^T\}$$

Thus we get that

$$y \sim \mathcal{N}_n[X\beta, \sigma^2\{I_n + ZQ(\psi)Z^T\}].$$

We can use the following simpler notation

$$\{I_n + ZQ(\psi)Z^T\} = \Upsilon^{-1}(\psi),$$

with ψ denoting the vector of distinct variance ratios appearing in Υ^{-1}. The model with the previous notations is

$$y \sim \mathcal{N}_n(X\beta, \sigma^2\Upsilon^{-1}(\psi)).$$

Then for known ψ the MLEs, classical results for weighted linear regression leads to

$$\widehat{\beta}_\psi = (X^{\mathrm{T}}\Upsilon X)^{-1}X^{\mathrm{T}}\Upsilon y, \quad \widehat{\sigma}^2_\psi = n^{-1}(y - X\widehat{\beta})^{\mathrm{T}}\Upsilon(y - X\widehat{\beta}).$$

b) **[Seen/unseen, 5 points]** The likelihood of the model is

$$\ell(\beta, \sigma^2, \psi) \equiv -\frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}\Upsilon(y - X\beta) - \frac{n}{2}\log\sigma^2 + \tfrac{1}{2}\log|\Upsilon|,$$

In the normal mixed model we take $\beta \equiv \lambda$ and note that if all the variance parameters are fixed, then $s_\psi = \widehat{\beta}_\psi = (X^{\mathrm{T}}\Upsilon X)^{-1}X^{\mathrm{T}}\Upsilon y$ is sufficient for β; its distribution is $\mathcal{N}_p\{\beta, \sigma^2(X^{\mathrm{T}}\Upsilon X)^{-1}\}$. Apart from constants, the logarithm of the required conditional density is therefore

$$\begin{aligned}\log f(y;\psi,\beta,\sigma^2) - \log f(\widehat{\beta}_\psi;\psi,\beta,\sigma^2) \quad &\equiv \quad -\frac{n}{2}\log\sigma^2 + \tfrac{1}{2}\log|\Upsilon| - \frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}\Upsilon(y - X\beta) \\ &\quad +\frac{p}{2}\log\sigma^2 - \tfrac{1}{2}\log|X^{\mathrm{T}}\Upsilon X| + \frac{1}{2\sigma^2}(\widehat{\beta}_\psi - \beta)^{\mathrm{T}}X^{\mathrm{T}}\Upsilon X(\widehat{\beta}_\psi - \beta),\end{aligned}$$

which reduces to the given form on writing $y - X\beta = (y - X\widehat{\beta}_\psi) + X(\widehat{\beta}_\psi - \beta)$ in the first quadratic term and expanding out, noting that

$$\begin{aligned}(\widehat{\beta}_\psi - \beta)^{\mathrm{T}}X^{\mathrm{T}}\Upsilon(y - X\widehat{\beta}_\psi) &= (\widehat{\beta}_\psi - \beta)^{\mathrm{T}}X^{\mathrm{T}}\Upsilon\{y - X(X^{\mathrm{T}}\Upsilon X)^{-1}X\Upsilon\} \\ &= (\widehat{\beta}_\psi - \beta)^{\mathrm{T}}\{X^{\mathrm{T}}\Upsilon y - X^{\mathrm{T}}\Upsilon X(X^{\mathrm{T}}\Upsilon X)^{-1}X\Upsilon y\} \\ &= 0.\end{aligned}$$

Using the previous calculation, we get

$$\log f(y;\psi,\beta,\sigma^2) - \log f(\widehat{\beta}_\psi;\psi,\beta,\sigma^2) \equiv \tfrac{1}{2}\log|\Upsilon| - \tfrac{1}{2}\log|X^{\mathrm{T}}\Upsilon X| - \frac{n-p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\widehat{\beta}_\psi)^{\mathrm{T}}\Upsilon(y - X\widehat{\beta}_\psi),$$

which proves that $\widehat{\beta}_\psi$ is a sufficient statistic for β and that we have the desired decomposition of the likelihood. And thus, the difference equals

$$\log f(y \mid \widehat{\beta}_\psi, \sigma^2, \psi) = \ell(\widehat{\beta}_\psi, \sigma^2, \psi) + \frac{p}{2}\log\sigma^2 - \tfrac{1}{2}\log|X^{\mathrm{T}}\Upsilon X|$$

i.e.,

$$\tfrac{1}{2}\log|\Upsilon| - \tfrac{1}{2}\log|X^{\mathrm{T}}\Upsilon X| - \frac{1}{2\sigma^2}(y - X\widehat{\beta}_\psi)^{\mathrm{T}}\Upsilon(y - X\widehat{\beta}_\psi) - \frac{n-p}{2}\log\sigma^2,$$

which allows to obtain estimators from the desired expressions.

c) **[Seen/unseen, 2 points]** We suppose now that Z is absent. Then

$$\log f(y;\widehat{\beta}_\psi, \sigma^2, \psi) = -\tfrac{1}{2}\log|X^{\mathrm{T}}X| - \frac{n-p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\widehat{\beta}_\psi)^{\mathrm{T}}(y - X\widehat{\beta}_\psi),$$

and taking partial derivative with regard to σ^2, we get

$$\widehat{\sigma}^2 = \tfrac{(y - X\widehat{\beta}_\psi)^{\mathrm{T}}(y - X\widehat{\beta}_\psi)}{n-p},$$

which differs from the classical likelihood estimator and is unbiased.

Exercise 5.

a) **[Seen, 3 points]** The Poisson model has no conditioning, so the log likelihood is

$$
\begin{aligned}
\ell_{\text{Poiss}}(\mu_1, \ldots, \mu_d) &= \sum_{i=1}^{d} y_i \log \mu_i - \mu_i - \log y_i! \\
&= m \log \left(\sum_{i=1}^{d} \mu_i \right) - \sum_{i=1}^{d} \mu_i - \log \left(\sum_{i=1}^{d} y_i \right)! + \log \tfrac{m!}{y_1! \ldots y_d!} + \sum_{i=1}^{d} y_j \log \tfrac{\mu_i}{\sum_{i=1}^{d} \mu_i}, \\
&= \ell_{\text{Poiss}}(m, \tau) + \ell_{\text{Multi}}(\beta, m, \pi_1, \ldots, \pi_d)
\end{aligned}
$$

where $\pi_i = \frac{x_i^T \beta}{\sum_{i=1}^{d} x_i^T \beta}$, $i = 1, \ldots, d$ and $\tau = \sum_{i=1}^{d} \mu_i$.

This result shows that we can fit a Poisson model with fixed sampling size of population to estimate the effect of the detergent.

b) **[Unseen, 4 points]** In the model, we want to estimate the preference, i.e., the `Brand` with a Poisson model and use the simplification above because sampling size is fixed within categories. That means that such a factorisation is possible, and this can be done with the model

$$
y \sim \gamma_{i,j,k} + \gamma_l,
$$

where $i = "low" or "High"$, $j = "B.User" or "Not.B.User$, $k = "Soft", "Medium" or "Hard"$ and $l = "A" or "B"$. This model is written

```
y \vim B.User * Temp * Soft + Brand
```

c) **[Unseen, 3 points]** From the `R` output, we see that adding `B.User * Brand` significantly descreases the deviance, revealing an effet of the variable `B.User` on the preference of the brand. Similar but less obvious effect appear when adding a category for `Temp`. No significant improvement is gained by adding categories for `Soft`. Finally the deviance being around 8 for the third model is a quite good fit.