Sketch of solutions for Homework 3

1. *Profile log-likelihood.* Suppose $Y = (Y_1, \ldots, Y_n)$ is a vector of independent, identically distributed random variables from the density $f(y; \psi, \lambda)$, where $\psi \in \mathbb{R}$ is the parameter of interest and $\lambda \in \mathbb{R}$ is a nuisance parameter. The profile log-likelihood is defined as $\ell_{\mathrm{p}}(\psi) = \ell(\psi, \hat{\lambda}_\psi)$, where $\hat{\lambda}_\psi$ is assumed to satisfy the score equation $\partial \ell(\psi, \lambda)/\partial \lambda = 0$.

   (a) Show that the estimator of ψ that satisfies the profile score equation $\partial \ell_{\mathrm{p}}(\psi)/\partial \psi = 0$ is the same as the maximum likelihood estimator of ψ.

   It's necessary to be a little bit careful with notation.

$$\ell'_{\mathrm{p}}(\psi) = \frac{\partial}{\partial \psi}\ell(\psi, \hat{\lambda}_\psi) + \frac{\partial}{\partial \lambda}\ell(\psi, \hat{\lambda}_\psi)\hat{\lambda}_\psi = \frac{\partial}{\partial \psi}\ell(\psi, \hat{\lambda}_\psi), \tag{1}$$

   by the assumption that $\hat{\lambda}_\psi$ satisfies the score equation. Both $\ell_{\mathrm{p}}(\psi)$ and $\hat{\lambda}_\psi$ are functions of y as well as ψ, but the dependence on y is suppressed, so $\ell'_p(\psi)$ means differentiation with respect to ψ. Denote the solution of (1) by $\hat{\psi}_{\mathrm{p}}$. Then

$$\frac{\partial}{\partial \psi}\ell(\hat{\psi}_{\mathrm{p}}, \hat{\lambda}_{\hat{\psi}_{\mathrm{p}}}) = 0.$$

   Also

$$\frac{\partial}{\partial \lambda}\ell(\hat{\psi}_{\mathrm{p}}, \hat{\lambda}_{\hat{\psi}_{\mathrm{p}}}) = 0,$$

   because the score equation is defined for each value of ψ, including $\hat{\psi}_{\mathrm{p}}$. The maximum likelihood estimators $\hat{\psi}$ and $\hat{\lambda}$ of ψ and λ are the simultaneous solution to

$$\frac{\partial}{\partial \psi}\ell(\hat{\psi}, \hat{\lambda}) = 0, \frac{\partial}{\partial \lambda}\ell(\hat{\psi}\hat{\lambda}) = 0. \tag{2}$$

   We assume that this solution is unique and is the maximum likelihood estimator, which implies $(\hat{\psi}_{\mathrm{p}}, \hat{\lambda}_{\hat{\psi}_{\mathrm{p}}}) = (\hat{\psi}, \hat{\lambda})$.

   I'm being very careful here, because many people lost marks for careless notation, and mixing up derivatives of ℓ_{p} with derivatives of $\ell(\psi, \lambda)$.

   (b) Show that the profile information function $j_{\mathrm{p}}(\psi) = -\partial \ell_{\mathrm{p}}(\psi)/\partial \psi \partial \psi^T$ satisfies

$$\{j_{\mathrm{p}}(\psi)\}^{-1} = j^{\psi\psi}(\psi, \hat{\lambda}_\psi),$$

   where $j^{\psi\psi}(\theta)$ is the (ψ, ψ) block of $j^{-1}(\theta)$, the inverse of the observed Fisher information from the log-likelihood function $\ell(\psi, \lambda)$.

Most people got this part, more or less carefully. First apply the chain rule to the derivative in (1), to get

$$j_{\rm p}(\psi) = j_{\psi\psi}(\psi, \hat\lambda_\psi) + j_{\psi\lambda}(\psi, \hat\lambda_\psi)\hat\lambda'_\psi,$$

and then use the score equation defining $\hat\lambda_\psi$ to show that $\hat\lambda'_\psi = -j_{\psi\lambda}(\psi, \hat\lambda_\psi)/j_{\lambda\lambda}(\psi, \hat\lambda_\psi)$. These combine to give

$$j_{\rm p}(\psi) = j_{\psi\psi}(\psi, \hat\lambda_\psi) - j_{\psi\lambda}(\psi, \hat\lambda_\psi) j_{\lambda\lambda}(\psi, \hat\lambda_\psi)^{-1} j_{\lambda\psi}(\psi, \hat\lambda_\psi).$$

Then use the formula for the inverse of a partitioned matrix to get the result. In general terms, this formula is

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & . \\ . & . \end{pmatrix}.$$

The ShermanMorrisonWoodbury formula gives another expression for the element in the RHS:

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1},$$

where if A is $n \times n$, B is $n \times p$, C is $p \times n$ and D is $p \times p$, with $p << n$, this can be useful if you already have A^{-1}, and can express your matrix of interest as a low rank adjustment to A. This seems to come up a lot in machine learning.

(c) Use Taylor series expansion to show that

$$\hat\lambda_\psi - \hat\lambda = -j^{-1}_{\lambda\lambda}(\hat\psi, \hat\lambda) j_{\lambda\psi}(\hat\psi, \hat\lambda)(\psi - \hat\psi) + O_p(n^{-1}).$$

There are two equally good ways to solve this. One is to consider $\hat\lambda_\psi$ as a function of ψ, and expand it about $\hat\psi$:

$$\hat\lambda_\psi = \hat\lambda_{\hat\psi} + (\psi - \hat\psi)\, d\hat\lambda_\psi/d\psi\Big|_{\hat\psi} + R,$$

and note that in (a) and (b) we established $\hat\lambda_{\hat\psi} = \hat\lambda$, and $d\hat\lambda_\psi/d\psi = -j_{\psi\lambda}(\psi, \hat\lambda_\psi)/j_{\lambda\lambda}(\psi, \hat\lambda_\psi)$. Then it just remains to show $R = O_p(n^{-1})$. In this version $R = (1/2)(\psi - \hat\psi)^2\, \hat\lambda''_\psi\Big|_{\psi^*}$, and work is needed to verify that $\hat\lambda'' = O_p(1)$; from work in class we have $(\psi - \hat\psi)^2 = O_p(n^{-1})$.

The other expansion is

$$\begin{aligned} 0 &= \frac{\partial}{\partial\lambda}\ell(\psi, \hat\lambda_\psi) \\ &= \frac{\partial}{\partial\lambda}\ell(\hat\psi, \hat\lambda) + (\psi - \hat\psi)\frac{\partial^2}{\partial\lambda\partial\psi}\ell(\hat\psi, \hat\lambda) + (\hat\lambda_\psi - \hat\lambda)\frac{\partial^2}{\partial\lambda^2}\ell(\hat\psi, \hat\lambda) + R, \end{aligned}$$

where we have used again that $\hat{\lambda} = \hat{\lambda}_{\hat{\psi}}$. Then we have

$$\begin{aligned}(\hat{\lambda}_\psi - \hat{\lambda}) j_{\lambda\lambda}(\hat{\theta}) + R &= (\hat{\psi} - \psi) j_{\psi\lambda}(\hat{\theta}), \\ \hat{\lambda}_\psi - \hat{\lambda} &= (\hat{\psi} - \psi) j_{\psi\lambda}(\hat{\theta}) j_{\lambda\lambda}^{-1}(\hat{\theta}) \{1 + R j_{\lambda\lambda}^{-1}(\hat{\theta})\}^{-1}\end{aligned}$$

where I'm assuming both λ and ψ are scalars, but the expressions hold for vectors as well. Now R will involve various third derivatives of ℓ, multiplied by either $(\psi - \hat{\psi})^2$, $(\hat{\lambda}_\psi - \hat{\lambda})^2$, or $(\psi - \hat{\psi})(\hat{\lambda}_\psi - \hat{\lambda})$. Each of the 3rd derivatives will be $O_p(n)$ by assumption, and $\hat{j}_{\lambda\lambda}^{-1}$ is $O_p(n^{-1})$. So we need $(\psi - \hat{\psi})^2$, $(\hat{\lambda}_\psi - \hat{\lambda})^2$, and $(\psi - \hat{\psi})(\hat{\lambda}_\psi - \hat{\lambda})$ to be $O_p(n^{-1})$. This is true for $\psi - \hat{\psi}$ by the asymptotic normality result, so we only need $\hat{\lambda}_\psi - \hat{\lambda} = O_p(n^{-1/2})$, which follows from the leading term of the expression above, or, by showing that $\hat{\lambda}_\psi = \lambda + O_p(n^{-1/2})$ from the score equation with ψ fixed, and $\hat{\lambda} = \lambda + O_p(n^{-1/2})$ from the pair of score equations for the full maximum likelihood estimator.

(d) Expand $\ell_{\mathrm{p}}(\psi)$ about $\hat{\psi}$ and use the results of (b) and (c) to show that

$$w_{\mathrm{p}}(\psi) = 2\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\} = (\psi - \hat{\psi})^2 j_{\mathrm{p}}(\hat{\psi}) + o_p(1),$$

and hence that the limiting distribution of $w_{\mathrm{p}}(\psi)$ is χ^2_1, under the model.

It's easy to get the first part from a Taylor series expansion in only ψ, modulo showing that $\ell'''_{\mathrm{p}}(\hat{\psi}) = o_p(1)$. This last step involves writing this 3rd derivative as a lengthy combination of 3rd derivatives of the original log-likelihood function, multiplied by terms like $(\psi - \hat{\psi})^3$, $(\psi - \hat{\psi})^2(\hat{\lambda}_\psi - \hat{\lambda})$, and so on, and these are $O_p(n^{-3/2})$. I was very unfussy about you checking the error term in this part, if you had done it reasonably carefully in (c).

Many people did not finish by showing the limiting distribution to be χ^2_1. It's not completely trivial, because $(\hat{\psi} - \psi) \dot{\sim} N(0, j^{\psi\psi}(\hat{\theta}))$, so $(\hat{\psi} - \psi)^2 \{j^{\psi\psi}(\hat{\theta})\}^{-1} \dot{\sim} \chi^2_1$; then use (b).

2. *BNC, Exercise 3.6.* Based on observations $y_1, \ldots y_n$ independently normally distributed with unknown mean and variance, obtain the profile log-likelihood for $\Pr(Y > a)$, where a is an arbitrary constant, and compare inference based on this with the exact answer from the non-central t-distribution.

With thanks to Evgeny Levi, Zhenhua Lin and Victor Veitch. First $\psi = \Phi((a - \mu)/\sigma)$, so $\mu = \sigma\Phi^{-1}(\psi) + a$, and expressions are neater if we let $\delta = \Phi^{-1}(\psi)$ and convert back at the end. Expressions are also simpler if we write $x_i = y_i - a$. The log-likelihood function is then

$$\ell(\delta, \sigma) = -n \log \sigma - (1/2\sigma^2) \sum (x_i - \sigma\delta)^2$$

and

$$\frac{\partial \ell(\delta, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum (x_i - \sigma\delta)^2 + \frac{1}{\sigma^2} \sum (x_i - \sigma\delta)\delta,$$

and setting this $= 0$ leads to a quadratic equation in σ_δ

$$-n\sigma^2 + ns_x - n\delta\bar{x}\sigma = 0, \qquad \text{where } s_x = \Sigma x_i^2/n, \bar{x} = \Sigma x_i/n$$

with solution

$$\hat{\sigma}_\delta = -\frac{\delta\bar{x}}{2} + \sqrt{s_x + \frac{\delta^2\bar{x}^2}{4}}.$$

We take the positive square root to ensure $\hat{\sigma}_\delta > 0$. From this the Wald pivot $(\delta - \hat{\delta})j_{\mathrm{p}}^{1/2}(\hat{\delta})$, or alternatively $(\psi - \hat{\psi})j_{\mathrm{p}}^{1/2}(\hat{\psi})$, can be computed numerically, although these will not lead to the same confidence intervals. The invariant pivot $r = \pm 2\{\ell_{\mathrm{p}}(\hat{\delta}) - \ell_{\mathrm{p}}(\delta)\}$ can be used if preferred.

It doesn't seem possible to make further progress except numerically. Lower bounds, or two-sided intervals, can be obtained from inverting r, or from $\hat{\psi} - z_\alpha j_{\mathrm{p}}^{-1/2}(\hat{\psi})$.

The bounds based on the non-central t distribution are given by $\Phi(T_{n-1}^{-1}\{(\alpha, n-1, \sqrt{n}((\mu - a)/\sigma)/\sqrt{n-1}\}$, where $T_{n-1}^{-1}(\cdot)$ is the quantile function (`qt`) for the non central T. However, I don't see how to get this without knowing μ, σ. So I think if the likelihood result is to be compared to this on the basis of a simulated sample of size n, that rather than use the true values μ and σ that generated the simulation, one should use the sample mean and sample variance in the non centrality parameter.

This problem is one of a class of 'non-standard' normal theory problems. The parameter ψ is the probability of an 'extreme result' $(Y > a)$, and it is often of interest to bound the confidence limit on such a probability. The parameter δ is sometimes called a tolerance limit. Similar calculations arise in regression calibration, where the parameter of interest is the value of x for which $y = y^*$, some specified value, in a linear model with $y = \beta_0 + \beta_1 x$, for example.

3. *Adapted from BNC, Ex. 2.24.*

   (a) Suppose $Y_1, \ldots, Y_n$ are independent, identically distributed as Poisson with mean θ. Show that the conditional distribution of $Y_1, \ldots, Y_n$, given $S = \Sigma Y_i$, is Multinomial(S, π) where $\pi = (1/n, \ldots, 1/n)$.
   This distribution can in principle be used to assess goodness of fit of the Poisson model, but if n is much bigger than 2 or 3 it will be difficult to determine which directions in the sample space to examine.

   (b) A summary statistic that could be used to see whether data are consistent with the moment properties of the Poisson is $T = \Sigma(Y_i - \bar{Y})^2/\{(n-1)\bar{Y}\}$. Show that

$$\mathrm{E}(T \mid S = s) = 1, \quad \mathrm{var}(T \mid S = s) = \frac{2(1 - 1/s)}{n-1},$$

   and thus that, conditionally on $S = s$, $(n-1)sT/(s-1)$ has the same first two moments as a $\chi^2_{(n-1)s/(s-1)}$.

This calculation was somewhat more brutal than I intended. Using the trinomial distribution for (Y_i, Y_j) gives a slightly easier form of the moment generating function. Even better is to note that the conditional distribution of Y_i, given Y_j and S, is binomial, with suitably defined parameters. It is fine with me if you use Wolfram alpha or Matlab or whatever for these calculations, but you should state this clearly. A similar comment applies to Q4.

(c) Explore the extension of this to assessing goodness of fit for a Poisson regression, where $y_i \sim \mathrm{Po}(\theta_i)$, and $\log \theta_i = \alpha + \beta x_i$.

I didn't mark this part, there doesn't seem to be anything very nice. Because it's an exponential family y, given Σy_i and $\Sigma y_i x_i$, is free of α and *beta*, but it's not clear what statistic to use to summarize this. In AS II, you may have learned to use the deviance as a goodness-of-fit statistics, treating it as approximately χ^2_{n-1}, and this is acceptable, but doesn't involve any conditioning as far as I know.

4. *SM, Problem 4.9.1.* The logistic density is a location-scale family with density function

$$f(y; \mu, \sigma) = \frac{\exp\{(y-\mu)/\sigma\}}{\sigma[1 + \exp\{(y-\mu)/\sigma\}]}, \quad -\infty < y < \infty, -\infty < \mu < \infty, \sigma > 0.$$

(a) When $\sigma = 1$, show that the expected Fisher information about μ in y is $1/3$.

(b) If instead of observing y, we observe $z = 1$ if $y > 0$, otherwise $z = 0$. When $\sigma = 1$ show that the maximum expected Fisher information about μ in z is $1/4$, achieved at $\mu = 0$, so that the maximum relative efficiency is $3/4$.

The point of this (easy) question is to show the loss in information by dichotomizing – it is roughly equivalent to throwing away at least 1/4 of the observations.

5. *Saddlepoint approximation.* Suppose $X_1, \ldots, X_n$ are independent and identically distributed on $\mathbb{R}$, with density function $f(x)$ and moment generating function $M_X(t) = E\{\exp(tX)\}$ assumed to exist for t in an open interval about 0, and cumulant generating function $K_X(t) = \log M_X(t)$. The *saddlepoint approximation* to the density of $\bar{X} = n^{-1}\Sigma X_i$ is given by

$$f_{\bar{X}}(\bar{x}) \doteq \frac{1}{\sqrt{2\pi}} \left\{ \frac{n}{K''_X(\hat{\phi})} \right\}^{1/2} \exp\{nK_X(\hat{\phi}) - n\hat{\phi}\bar{x}\},$$

where $\hat{\phi} = \hat{\phi}(\bar{x})$ satisfies the equation $K'_X(\hat{\phi}) = \bar{x}$.

(a) Show that if $Y_1, \ldots, Y_n$ are independent and identically distributed from a scalar parameter exponential family

$$f(y;\theta) = \exp\{\theta y - c(\theta) - d(y)\}$$

that the saddlepoint approximation to the density of $\hat{\theta}$ is given by

$$f_{\hat{\Theta}}(\hat{\theta};\theta) \doteq \frac{1}{\sqrt{2\pi}} j^{1/2}(\hat{\theta}) \exp\{\ell(\theta) - \ell(\hat{\theta})\}.$$

(b) If $y_1, \ldots, y_n$ are independent and identically distributed from a scalar parameter location family

$$f(y;\theta) = f_0(y - \theta),$$

then we showed in class that the exact density of the maximum likelihood estimator $\hat{\theta}$, given a, where $a_i = y_i - \hat{\theta}, i = 1, \ldots, n$, is

$$f_{\hat{\Theta}|A}(\hat{\theta} \mid a;\theta) = \frac{\exp\{\ell(\theta;y)}{\int \exp\{\ell(\theta;y)\} d\theta},$$

where in the right hand side we recall that $y_i = \hat{\theta} + a_i$. By expanding $\ell(\theta)$ in the denominator in a Taylor series about $\hat{\theta}$, show that the exact conditional density can be approximated by

$$f_{\hat{\Theta}|A}(\hat{\theta} \mid a;\theta) \doteq \frac{1}{\sqrt{2\pi}} j^{1/2}(\hat{\theta}) \exp\{\ell(\theta) - \ell(\hat{\theta})\}.$$

Both these approximations have similar versions for p-dimensional parametric models, with slight changes in notation. Both approximations have relative error $O(n^{-1})$, and when re-normalized to integrate to 1 have relative error $O(n^{-3/2})$.

[Not everyone finished (a), by making the (1-1) transformation from $\bar{y}$ to $\hat{\theta}$, which moves the $j(\hat{\theta})^{1/2}$ from the denominator to the numerator.]

The $O(n^{-1})$ error in the saddle point approximation comes from the saddlepoint expansion, but expecting you to prove this would be unfair in the extreme. You would need to know that

$$f_{\bar{X}}(\bar{x}) = \frac{1}{\sqrt{2\pi}} \left\{ \frac{n}{K_X''(\hat{\phi})} \right\}^{1/2} \exp\{nK_X(\hat{\phi}) - n\hat{\phi}\bar{x}\}\{1 + \frac{d(\hat{\phi})}{n} + O_p(\frac{1}{n^2})\},$$

where $d(\hat{\phi})$ involves 3rd and 4th standardized derivatives of $K_x(\phi)$. In fact it is

$$\frac{1}{24}\left(-\frac{\{3K_X'''(\hat{\phi})\}^2}{K_X''(\hat{\phi})^{3/2}} + \frac{5K_X^{(4)}(\hat{\phi})}{K_X''(\hat{\phi})^2} \right),$$

(Reid, *Statistical Science*, 1988, p.213-227) This then gives

$$f_{\hat{\Theta}}(\hat{\theta};\theta) = \frac{1}{\sqrt{2\pi}} j^{1/2}(\hat{\theta}) \exp\{\ell(\theta) - \ell(\hat{\theta})\}\{1 + \frac{\tilde{d}(\hat{\theta})}{n} + O(\frac{1}{n^2})\}.$$

where $\tilde{d}(\hat{\theta})$ is a similar function of standardized 3rd and 4th derivatives of $\ell(\theta)$, evaluated at $\hat{\theta}$. We could then write $\tilde{d}(\hat{\theta}) = \tilde{d}(\theta)\{1 + O_(n^{-1/2})\}$, and the $\tilde{d}(\theta)$ term would be absorbed into the renormalizing constant giving

$$f_{\hat{\Theta}}(\hat{\theta};\theta) = c j^{1/2}(\hat{\theta}) \exp\{\ell(\theta) - \ell(\hat{\theta})\}\{1 + O_(n^{-3/2})\}.$$

Don't worry this will *NOT* be on the test. To get it rigorous involves even more work, because this substitution only works for $\hat{\theta}$ in an $O(n^{-1/2})$ neighbourhood of θ, so we have to show that we can ignore the contribution from outside this neighbourhood.

The error term in the Laplace approximation is easier. For example, expand the integral in the denominator to fourth order, to get

$$\ell(\theta) \doteq \ell(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 \ell''(\hat{\theta}) + \frac{1}{6}(\theta - \hat{\theta})^3 \ell'''(\hat{\theta}) + \frac{1}{24}(\theta - \hat{\theta})^4 \ell^{(4)}(\hat{\theta});$$

now when we integrate $\exp \ell(\theta)$ with respect to θ we have:

$$\begin{aligned} & \int \exp[\frac{1}{2}(\theta - \hat{\theta})^2 \ell''(\hat{\theta})\{1 + \frac{1}{3}(\theta - \hat{\theta})^3 \ell'''(\hat{\theta})/\ell''(\hat{\theta}) + \frac{1}{12}(\theta - \hat{\theta}^2)\ell^{(4)}(\hat{\theta})/\ell''(\hat{\theta})\}]d\theta \\ = & \int e^{-\frac{1}{2}(\theta-\hat{\theta})^2 j(\hat{\theta})}[1 + \frac{1}{3}(\theta - \hat{\theta})^3 \ell'''(\hat{\theta})/\ell''(\hat{\theta}) + \frac{1}{12}(\theta - \hat{\theta}^2)\ell^{(4)}(\hat{\theta})/\ell''(\hat{\theta})\} \\ & + \frac{1}{18}(\theta - \hat{\theta})^6 \{\ell'''(\hat{\theta})/\ell''(\hat{\theta})\}^2]d\theta \end{aligned}$$

and the remainder terms are the central moments for θ in the $N(\hat{\theta}, j^{-1}(\hat{\theta}))$ distribution.