ECOM40006/90013 Econometrics 3

Week 10 Extras (Part 3 Solutions)

Question 1: The Story So Far (Normal Distribution Revisited)

Note. Don't take the level of working here to be indicative of what you'd actually do on an exam. The goal is to carefully step out exactly what is being done in some of the more troublesome lines, but there also runs the risk of the explanation being a bit too heavy. If you get the working, great! If it's too dense or there are things that aren't clear, let me know.

(a) Notation. Define $\theta \equiv (\beta, \sigma^2)$. The log-likelihood is

$$\begin{aligned}
\log L_n(\theta) &= \sum_{i=1}^{n} \log f(y_i; x_i, \theta) \\
&= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}\right) \\
&= \sum_{i=1}^{n} \log \left[(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}\right)\right] \\
&= \sum_{i=1}^{n} \left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{(y_i - x_i'\beta)^2}{2\sigma^2}\right] \\
&= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2
\end{aligned}$$

(b) The score requires that we take two partial derivatives of the log-likelihood: one for β and one for σ^2. The partial with respect to β is

$$\begin{aligned}
\frac{\partial \log L_n(\theta)}{\partial \beta} &= \frac{\partial}{\partial \beta}\left(-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right) \\
&= \frac{\partial}{\partial \beta}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right) \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial (y_i - x_i'\beta)^2}{\partial \beta}
\end{aligned}$$

where the second equality follows from the fact that the first few terms do not have β in them at all (and hence differentiates out), and the last equality from the fact that the derivative of a sum is the sum of the derivatives (e.g. the partial of $a + b$ is the partial

of a plus the partial of b, of which this is extended to a sum of n terms). This means it suffices to find

$$\begin{aligned}\frac{\partial(y_i - x_i'\beta)^2}{\partial\beta} &= 2(y_i - x_i'\beta)\left(-\frac{\partial x_i'\beta}{\partial\beta}\right) \\ &= -2x_i(y_i - x_i'\beta)\end{aligned}$$

using the (hopefully by now) familiar result that $\frac{\partial x_i'\beta}{\partial\beta} = x_i$.[1] Therefore, the first partial of note is

$$\begin{aligned}\frac{\partial \log L_n(\theta)}{\partial\beta} &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}[-2x_i(y_i - x_i'\beta)] \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i y_i - x_i x_i'\beta).\end{aligned}$$

The second partial is then, eliminating (additive) terms that do not have σ^2 in them:

$$\begin{aligned}\frac{\partial \log L_n(\theta)}{\partial\sigma} &= \frac{\partial}{\partial\sigma^2}\left[-\frac{n}{2}\log\sigma^2 - \frac{1}{2}\sigma^{-1}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right] \\ &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\end{aligned}$$

Therefore, the score is

$$\begin{aligned}s_n(\theta) = \frac{\partial \log L_n(\theta)}{\partial(\theta)} &= \begin{bmatrix}\dfrac{\partial \log L_n(\theta)}{\partial\beta} \\ \dfrac{\partial \log L_n(\theta)}{\partial\sigma^2}\end{bmatrix} \\ &= \begin{bmatrix}\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i y_i - x_i x_i'\beta) \\ -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(y_i - x_i'\beta)^2\end{bmatrix}\end{aligned}$$

Solving for the respective first-order conditions allows us to solve for the maximum likelihood estimators (MLEs) of the problem. For example, at $\hat{\theta} = (\hat{\beta}, \hat{\sigma}^2)$, the first partition

[1] Some of you may note that $y_i - x_i'\beta$ got moved to the right hand side. This is a valid operation because $y_i - x_i'\beta$ is a scalar, but you may notice that βx_i is not a conformable matrix for multiplication. Moving $y_i - x_i'\beta$ to the right solves this problem.

of $s_n(\hat{\theta})$ lets us solve in this case for $\hat{\beta}$ directly:

$$\begin{aligned}
\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i y_i - x_i x_i'\hat{\beta}) &= 0 \\
\implies \sum_{i=1}^{n} x_i x_i'\hat{\beta} &= \sum_{i=1}^{n} x_i y_i \\
\implies \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\sum_{i=1}^{n} x_i x_i'\right)\hat{\beta} &= \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1}\sum_{i=1}^{n} x_i y_i \\
\implies \hat{\beta} &= \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1}\sum_{i=1}^{n} x_i y_i
\end{aligned}$$

Note that this is the usual OLS estimator. The second partition satisfies

$$\begin{aligned}
-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 &= 0 \\
\implies \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 &= \frac{n}{2\hat{\sigma}^2} \\
\implies \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2
\end{aligned}$$

so that the MLEs are

$$\begin{aligned}
\hat{\beta} &= \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1}\sum_{i=1}^{n} x_i y_i \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2.
\end{aligned}$$

(c) To check whether or not the MLEs found in part (b) are indeed maximizers in the sense that they maximize the log-likelihood, we'll need to check the Hessian. Specifically, we want to see whether or not the Hessian is *negative definite* at $\hat{\theta}$. This can be thought of as the steps for a matrix version of the second derivative test, where in a univariate situation we'd check to see if a solution to the FOC is a maximizer by seeing whether or not the second derivative was negative at that solution.

The Hessian in this case will have four main components.

- The first two components we can find by taking the gradient of $\frac{\partial \log L_n(\cdot)}{\partial \beta}$ again.
- Similarly, doing the same thing for $\frac{\partial \log L_n(\cdot)}{\partial \sigma^2}$ gives the next two terms of the Hessian.

To ensure that the Hessian itself is square, we'll be taking the gradient with respect to the transposes. Fortunately, because of how these gradients work, an informal way to deal with this is to take the gradient as usual (i.e. pretend it doesn't have the transpose), then

transpose the final answer. To see this, let's look at the gradient of $\log L_n(\cdot)$ with respect to β. Call this (1) so that

$$(1) = \frac{\partial \log L_n(\theta)}{\partial \beta}$$

The two partials we can take here are with respect to β and with σ so that

$$\begin{aligned}
\frac{\partial(1)}{\partial\beta} &= \frac{\partial}{\partial\beta}\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_iy_i - x_ix_i'\beta) \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{\partial(x_iy_i - x_ix_i'\beta)}{\partial\beta} \\
&= -\frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{\partial x_ix_i'\beta}{\partial\beta} \\
&= -\frac{1}{\sigma^2}\sum_{i=1}^{n}x_ix_i'
\end{aligned}$$

since the matrix x_ix_i' is symmetric.[2]

If we take the partial with respect to σ^2 we get

$$\begin{aligned}
\frac{\partial(1)}{\partial\sigma^2} &= \frac{\partial}{\partial\sigma^2}\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_iy_i - x_ix_i'\beta) \\
&= \frac{\partial(\sigma^2)^{-1}}{\partial\sigma^2}\sum_{i=1}^{n}(x_iy_i - x_ix_i'\beta) \\
&= -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_iy_i - x_ix_i'\beta).
\end{aligned}$$

Continue to keep in mind that the variable of interest here is σ^2, not σ.

Therefore, the gradient vector here is

$$\frac{\partial(1)}{\partial\theta} = \begin{bmatrix}\frac{\partial(1)}{\partial\beta} \\ \frac{\partial(1)}{\partial\sigma^2}\end{bmatrix} = \begin{bmatrix}-\frac{1}{\sigma^2}\sum_{i=1}^{n}x_ix_i' \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_iy_i - x_ix_i'\beta)\end{bmatrix}$$

Notice that if we transpose this, we get

$$\frac{\partial(1)}{\partial\theta'} = \begin{bmatrix}-\frac{1}{\sigma^2}\sum_{i=1}^{n}x_ix_i' & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_iy_i - x_ix_i'\beta)\end{bmatrix}$$

[2]Why? Because it equals its own transpose. If you use the rule $(AB)' = B'A'$ you get $(x_ix_i')' = (x_i')'x_i' = x_ix_i'$. The hope is that you're used to this by now, but if not it never hurts to practice some more!

This is the first row of the Hessian matrix. Note that if one just wanted the Hessian itself, there's absolutely no need to do it this way, but this particular way might help for those who have difficulty understanding.

How about the second row? Well, let's define (2) to be

$$(2) = \frac{\partial \log L_n(\theta)}{\partial \sigma^2}$$

so that

$$\begin{aligned}
\frac{\partial(2)}{\partial\beta} &= \frac{\partial}{\partial\beta}\left(-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right) \\
&= \frac{\partial}{\partial\beta}\left(\frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right) \\
&= \frac{1}{2\sigma^4}\sum_{i=1}^{n}\frac{\partial(y_i - x_i'\beta)^2}{\partial\beta} \\
&= \frac{1}{2\sigma^4}\sum_{i=1}^{n}(-2x_i(y_i - x_i'\beta)) \\
&= -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i y_i - x_i x_i'\beta)
\end{aligned}$$

Notice that this is exactly the same as $\frac{\partial(1)}{\partial\sigma^2}$. This is not a coincidence, and is a feature of Hessian matrices that you can rely on to check your answers whenever necessary.[3] Furthermore,

$$\begin{aligned}
\frac{\partial(2)}{\partial\sigma^2} &= -\frac{n}{2}\frac{\partial(\sigma^2)^{-1}}{\partial\sigma^2} + \frac{1}{2}\frac{\partial(\sigma^2)^{-2}}{\partial\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \\
&= \frac{n}{2\sigma^4} - \frac{1}{2}\times 2\times(\sigma^2)^{-3}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \\
&= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - x_i'\beta)^2
\end{aligned}$$

Therefore, the gradient vector for (2) is

$$\frac{\partial(2)}{\partial\theta} = \begin{bmatrix}\frac{\partial(2)}{\partial\beta} \\ \frac{\partial(2)}{\partial\sigma^2}\end{bmatrix} = \begin{bmatrix} -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i y_i - x_i x_i'\beta) \\ \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\end{bmatrix}$$

[3] The exact result here is known as *Young's Theorem*. Provided that the problem you're dealing with is 'regular', the Hessian is symmetric so you don't have to solve for every single element in the Hessian in general. Of course, this theorem can fail, but this is not such an example.

The transpose of this is

$$\frac{\partial(2)}{\partial\theta'} = \left[-\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i y_i - x_i x_i'\beta) \quad \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right]$$

The overall Hessian. Putting all of these together, the Hessian can be written in full as

$$H_n(\theta) = \begin{bmatrix}\frac{\partial(1)}{\partial\theta'}\\ \frac{\partial(2)}{\partial\theta'}\end{bmatrix} = \begin{bmatrix} -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i x_i' & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i y_i - x_i x_i'\beta) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i y_i - x_i x_i'\beta) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \end{bmatrix}$$

(d) Following on from the intuition given in (c), now we need to check to see whether the Hessian is negative definite or not. From the FOCs, we know that the MLEs $\hat{\beta}$ and $\hat{\sigma}^2$ have to satisfy the property that

$$\sum_{i=1}^{n} x_i(y_i - x_i'\hat{\beta}) = 0$$

so that if we evaluate $H_n(\beta, \hat{\sigma}^2)$ at the MLEs, this immediately eliminates the off-diagonal elements, leaving us with

$$H_n(\hat{\theta}) = \begin{bmatrix} -\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n} x_i x_i' & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 \end{bmatrix}$$

Recall from the FOCs that we also have the result

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 \implies n\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2$$

so that the bottom right segment of the Hessian becomes

$$\begin{aligned}\frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 &= \frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6}n\hat{\sigma}^2 \\ &= \frac{n}{\hat{\sigma}^4}\left(\frac{1}{2} - 1\right) \\ &= -\frac{n}{\hat{\sigma}^4}\end{aligned}$$

Hence the Hessian reduces to

$$H_n(\hat{\theta}) = \begin{bmatrix} -\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n} x_i x_i' & 0 \\ 0 & -\frac{n}{\hat{\sigma}^4} \end{bmatrix}$$

Note: in the following working below, we're going to use something called 'partitioned' or 'block' matrix multiplication. If you just think of the Hessian as a 2×2 matrix for this example (which it isn't), you should be able to see what's going on here.

Observe that for any $(k+1) \times 1)$ vector $z \neq 0$ (this basically just means that whatever this vector z_i is, it doesn't have every element equal to zero. In other words, anything that's not the zero vector) one has

$$\begin{aligned} z' H_n(\hat{\theta}) z_i &= \begin{bmatrix} z_1' & z_2' \end{bmatrix} \begin{bmatrix} -\frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} x_i x_i' & 0 \\ 0 & -\frac{n}{\hat{\sigma}^4} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \\ &= -\frac{1}{\hat{\sigma}^2} z_1' \left(\sum_{i=1}^{n} x_i x_i' \right) z_1 - z_2' \frac{n}{\hat{\sigma}^4} z_2 \\ &< 0 \end{aligned}$$

due to the following observations:

- $\sum_{i=1}^{n} x_i x_i'$ is a positive definite matrix (if it wasn't, then it wouldn't be invertible and the MLE $\hat{\beta}$ wouldn't exist).
- Therefore, $z_1' \left(\sum_{i=1}^{n} x_i x_i' \right) z_1 > 0$ for any $z_1 \neq 0$, and the negative in front ensures that the whole expression is negative (it is understood that $\hat{\sigma}^2 > 0$ also).
- For similar reasons, $z_2' \frac{n}{\hat{\sigma}^4} z_2 > 0$ as long as $z_2 \neq 0$.
- While it is possible for one of z_1 or z_2 to be zero, it is not possible for both of them to be zero as $z \neq 0$ ensures that at least one of these terms is nonzero at all times.

Therefore, the Hessian is negative definite[4] at the MLEs, verifying that $\hat{\theta}$ does indeed maximize the log-likelihood.

(e) How does one conceptualize the score?

- Suppose that the true conditional mean is given by $\mathbb{E}(Y_i|X_i) = X_i'\beta_0$.
- In the working, we've used $X_i'\beta$ so far. That's because
  - We've assumed the same functional form in the working as $\mathbb{E}(Y_i|X_i)$
  - The score is a function of β

  If we, as econometricians, are fortunate enough to pick exactly the right functional form for whatever regression we're working with, the fact that the score depends on β (as well as σ^2) means that we can calculate this for any value of β and σ^2 we like.

  The point here then is that special things happen if we somehow manage to pick the *right* choice of parameters. Namely, $\beta = \beta_0$ and $\sigma^2 = \sigma_0^2$.

[4]The exact definition being used is that a symmetric matrix A is negative definite if for any conformable column vector $z \neq 0$,$z'Az < 0$.

Since we're working with population variables now, it would be technically more appropriate to re-express the score as

$$s_n(\theta) = \begin{bmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n} X_i(Y_i - X_i'\beta) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2 \end{bmatrix} = \begin{bmatrix} (1a) \\ (2a) \end{bmatrix}$$

Note that at the end of the day, what is more important than the attention to notation is being able to show an understanding of the underlying steps. In any case, we have

$$\mathbb{E}(s_n(\theta) = \begin{bmatrix} \mathbb{E}[(1a)] \\ \mathbb{E}[(2a)] \end{bmatrix}$$

where

$$\begin{aligned} \mathbb{E}[(1a)] &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbb{E}[X_i(Y_i - X_i'\beta)] \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbb{E}[\mathbb{E}(X_i(Y_i - X_i'\beta)|X_i)] && \text{(LIE)} \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbb{E}[X_i(\mathbb{E}(Y_i|X_i) - X_i'\beta)] \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbb{E}[X_i(X_i'\beta_0 - X_i'\beta] \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbb{E}[X_iX_i'(\beta_0 - \beta)]. \end{aligned}$$

In the third line, note that some extra properties of conditional expectations have been used.[5] Since this can be evaluated at any given β of our choosing, it turns out that this equals zero if we specifically evaluate this at $\theta = (\beta_0, \sigma_0^2)$. Note that the value of σ_0^2 doesn't really play much of a part here, but is here for completeness in particular.

Now for the next expectation. Since β and σ are treated as constants, we can use the linearity of conditional expectations directly skip to the line where we deal with the actual

[5] If you're having trouble following, what happened was that

$$\begin{aligned} \mathbb{E}(X_i(Y_i - X_i'\beta)|X_i) &= \mathbb{E}(X_iY_i - X_i'\beta|X_i) \\ &= \mathbb{E}(X_iY_i|X_i) - \mathbb{E}(X_i'\beta|X_i) && \text{(linearity, conditional expectations)} \\ &= X_i\mathbb{E}(Y_i|X_i) - X_i'\beta \end{aligned}$$

where we make use of the fact that functions of X_i are treated as constant when conditioned on X_i.

random variables themselves

$$\begin{aligned}
\mathbb{E}[(2a)] &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\mathbb{E}\left(\sum_{i=1}^{n}(Y_i - X_i'\beta)^2\right) \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\mathbb{E}[(Y_i - X_i'\beta)^2] \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\mathbb{E}[\mathbb{E}((Y_i - X_i'\beta)^2|X_i)] \qquad \text{(LIE)}
\end{aligned}$$

The trick here is to pre-emptively evaluate this at $\theta = (\beta_0, \sigma_0^2)$ so that we have

$$\begin{aligned}
\mathbb{E}[(2a)]\big|_{(\beta_0,\sigma_0^2)} &= -\frac{n}{2\sigma_0^2} + \frac{1}{2\sigma_0^4}\sum_{i=1}^{n}\mathbb{E}[\mathbb{E}((Y_i - X_i'\beta_0)^2|X_i)] \\
&= -\frac{n}{2\sigma_0^2} + \frac{1}{2\sigma_0^4}\sum_{i=1}^{n}\mathbb{E}[\mathbb{E}((Y_i - \mathbb{E}(Y_i|X_i))^2|X_i)] \\
&= -\frac{n}{2\sigma_0^2} + \frac{1}{2\sigma_0^4}\sum_{i=1}^{n}\mathbb{E}[\mathrm{var}(Y_i|X_i)] \qquad \text{(definition, conditional variance)} \\
&= -\frac{n}{2\sigma_0^2} + \frac{1}{2\sigma_0^4}\sum_{i=1}^{n}\sigma_0^2 \qquad \text{as } \mathbb{E}(\sigma_0^2) = \sigma_0^2 \\
&= -\frac{n}{2\sigma_0^2} + \frac{n\sigma_0^2}{2\sigma_0^4} \\
&= -\frac{n}{2\sigma_0^2} + \frac{n}{2\sigma_0^2} \\
&= 0.
\end{aligned}$$

Therefore

$$\mathbb{E}(s_n(\beta_0, \sigma_0^2)) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and so the expected score is zero at $(\beta, \sigma^2) = (\beta_0, \sigma_0^2)$.

Aside. It's also very common to see this kind of derivation being done for the score *for a single observation*. In this case we've decided to stay with the entire score because (i) we already have the score and (ii) it gives the same answer. After all, if the expected score for a single observation is zero, then the expected score for n observations is, thanks to the summation, n times the score for a single observation (which is still zero).