ECOM40006/90013 Econometrics 3

Week 11 Extras: Solutions

Question 1: GLM, Maximum Likelihood, and You

(a) (i.) The interpretation is as follows: *a one unit increase in x_i raises the probability that $y_i = 1$ by $\hat{\beta}_1$ on average, all else constant.* A fairly stock standard interpretation, with the only difference being that we are talking about changing the *probability* of $y_i = 1$ rather than y_i itself. This should hopefully make sense, since y_i can only ever take two values!

(ii.) The primary advantages of a linear probability model (LPM) are that it is (i) easy to estimate and (ii) easy to interpret, since standard OLS interpretations apply to a linear probability model!

(iii.) The drawbacks of the LPM can be quite severe. First and foremost, an LPM is capable of forecasting probabilities that are either (i) greater than 1 or (ii) less than zero. Both of these are nonsensical interpretations and the usefulness of the linear probability model can be dampened in these situations.

To make things worse, there's also no guarantee that OLS would be able to obtain the correct marginal coefficients $\beta_1, \beta_2, \ldots$ etc if the underlying *data generating process* (DGP) is latent. Namely, it's possible for the linear probability model to even give the wrong signs!

(b) If we set $y_i = 1$, we get

$$\mathbf{P}(y_i = 1) = \theta^1(1-\theta)^{1-1} = \theta.$$

If we set $y_i = 0$ instead, we find

$$\mathbf{P}(y_i = 0) = \theta^0(1-\theta)^{1-0} = 1-\theta.$$

So in effect, the probability mass function is a convenient way to represent the binary variable

$$y_i = \begin{cases} 1 & \text{with probability } \theta \\ 0 & \text{with probability } 1-\theta. \end{cases}$$

(c) The log-likelihood is

$$\begin{aligned}\log L(\theta) &= \sum_{i=1}^{n} \log p(y_i) \\ &= \sum_{i=1}^{n} \log \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \sum_{i=1}^{n} [y_i \log \theta + (1-y_i)\log(1-\theta)].\end{aligned}$$

We could go further, but because of some of the derivations we have to do later, it's more convenient to leave it like this for the time being. We then obtain the score by taking the gradient:

$$\begin{aligned}S(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} &= \sum_{i=1}^{n} \left[y_i \frac{\partial \log \theta}{\partial \theta} + (1-y_i)\frac{\partial \log(1-\theta)}{\partial \theta} \right] \\ &= \sum_{i=1}^{n} \left[\frac{y_i}{\theta} - \frac{1-y_i}{1-\theta} \right] \\ &= \sum_{i=1}^{n} \left(\frac{y_i(1-\theta) - \theta(1-y_i)}{\theta(1-\theta)} \right) \\ &= \sum_{i=1}^{n} \left(\frac{y_i - \theta y_i - \theta + \theta y_i}{\theta(1-\theta)} \right) \\ &= \sum_{i=1}^{n} \left(\frac{y_i - \theta}{\theta(1-\theta)} \right) \qquad (1)\end{aligned}$$

Keep this last equation in mind: this will come in very useful for when we start dealing with probit and logit models later (especially when it comes to calculating the generalized residual in the score for those types of models). From here, observe that the denominator $\theta(1-\theta)$ is constant, so it can be taken out of the summation. Note that this is the only time that we'll be able to do this; later on this won't be possible.

Taking the first-order condition (i.e. setting the score to zero) gives

$$\begin{aligned}S(\theta) = 0 &\implies \sum_{i=1}^{n} \left(\frac{y_i - \theta}{\theta(1-\theta)} \right) = 0 \\ &\implies \sum_{i=1}^{n} (y_i - \theta) = 0 \\ &\implies \sum_{i=1}^{n} y_i = n\theta \\ &\implies \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i.\end{aligned}$$

In other words, the MLE for the Bernoulli distribution is the sample mean. In other words, this can also be interpreted as the proportion of observations that equal 1 due to the role of y_i as a binary variable that only takes the values of 0 and 1.

(d) (i.) Models where y_i is generated via this kind of DGP are called *latent* variables models because the 'fitted value' that determines whether $y_i = 1$ is unobservable. Specifically, if we let $y_i^* = \beta_0 + \beta_1 x_i$, the value of y_i^* is unobservable.

Specifically: even though data on both y_i and x_i are observable, the parameters that implicitly generate y_i^* are not observable and so the actual value of y_i^* as a whole is not observable by the econometrician, even though x_i is known!

To put this in perspective, think of it like this:

- If $y_i^* = 1$, $y_i = 1$.
- If $y_i^* = 100$, $y_i = 1$.
- If $y_i^* = 9000$, $y_i = 1$.

So in fact, by not being able to observe y_i^*, there is a loss of information. This kind of conclusion also sheds some light on why the marginal effects are often not well approximated in a linear probability model.

(ii.) From the DGP highlighted at the start of the question, we know that $y_i = 1$ whenever

$$\begin{aligned} &\beta_0 + \beta_1 x_i + u_i > 0 \\ &\quad \implies u_i > -(\beta_0 + \beta_1 x_i). \end{aligned}$$

Because the event that $y_i = 1$ occurs whenever the event $u_i > -(\beta_0 + \beta_1 x_i)$ does, we have that, dependent on the value of x_i,

$$\begin{aligned} \mathbf{P}(y_i = 1|x_i) &= \mathbf{P}(u_i > -(\beta_0 + \beta_1 x_i)) \\ &= 1 - \mathbf{P}(u_i \leq -(\beta_0 + \beta_1 x_i)) \\ &= 1 - F(-(\beta_0 + \beta_1 x_i)) \\ &= F(\beta_0 + \beta_1 x_i), \end{aligned}$$

where $F(\cdot)$ represents the CDF of u_i. Note that the CDF here is always implicitly dependent on the value of x_i; for notational convenience we generally omit it (but is usually just lurking around in the background). The last line uses the fact that u_i is symmetric around zero. To see why, a picture would best illustrate this, namely,

$$1 - F(-x) = F(x).$$

One such picture that illustrates these symmetry properties would be the one that appears in question 1 of the week 7 extras.

(e) (i.) The main property of a CDF that makes it useful for predicting probabilities is the fact that its ending values are always between 0 and 1. In fact, the standard normal CDF here accepts *any* real number and returns a number in the interval [0,1]. Namely:

$$\Phi : \mathbb{R} \to [0, 1].$$

(*Note: so would it ever be the case that a function like this would ever return a value that is exactly zero or one? Not in this case. But the closed brackets are there just to remind you that we can also consider CDFs which are capable of that, with the CDFs of discrete random variables giving a good example of where that can crop up.*)

So effectively, this kind of model can be thought of as a generalized linear model in the sense that you can consider what we call the *linear predictors* $x_i'\beta$ from a linear model. If we have coefficient estimates $\hat{\beta}$, we can calculate the fitted values $x_i'\hat{\beta}$. Now, these fitted values could be outside the interval [0,1] by virtue of the fact that fitted values can be any real number. But this property that was a problem with the LPM isn't an issue anymore: the standard normal CDF transforms it into a probability that sits between 0 and 1.

(ii.) The *Chain Rule* is the main way to get these derivatives, noting that by the properties of the CDF

$$\frac{\partial \Phi(x)}{\partial x} = \phi(x),$$

where $\phi(x)$ is the PDF of the standard normal distribution (i.e. the usual bell curve). Then, we have

$$\frac{\partial \theta_i}{\partial \beta_0} = \frac{\partial \Phi(\beta_0 + \beta_1 x_i)}{\partial \beta_0 + \beta_1 x_i} \frac{\partial \beta_0 + \beta_1 x_i}{\partial \beta_0} = \phi(\beta_0 + \beta_1 x_i) \times 1 = \phi_i,$$

where $\phi_i = \phi(\beta_0 + \beta_1 x_i)$ as defined in the question. Similarly, the partial of θ_i with respect to β_1 is then

$$\frac{\partial \theta_i}{\partial \beta_1} = \frac{\partial \Phi(\beta_0 + \beta_1 x_i)}{\partial \beta_0 + \beta_1 x_i} \frac{\partial \beta_0 + \beta_1 x_i}{\partial \beta_1} = \phi(\beta_0 + \beta_1 x_i) \times x_i = \phi_i x_i.$$

(iii.) The log-likelihood in the question can actually be obtained from the log-likelihood in part (b)-(ii): all you have to do is replace θ by θ_i. This time, θ_i is no longer constant so it cannot move out of the summations. Before proceeding it would be a good idea to note the following: for $j = 0, 1$,

$$\frac{\partial \log \theta_i}{\partial \beta_j} = \frac{\partial \log \theta_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\theta_i} \frac{\partial \theta_i}{\partial \beta_j}$$

and

$$\frac{\partial \log(1-\theta_i)}{\partial \beta_j} = \frac{\partial \log(1-\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j} = -\frac{1}{1-\theta_i} \frac{\partial \theta_i}{\partial \beta_j}.$$

So in fact, when we take the score, we get a nearly identical expression to what we get in part (b)-(ii), but with an extra common factor $\frac{\partial \theta_i}{\partial \beta_j}$ that shows up at the front. To see this:

$$S(\beta) = \frac{\partial \log L(\theta)}{\partial \beta} = \begin{bmatrix} \dfrac{\partial \log L(\theta)}{\partial \beta_0} \\ \dfrac{\partial \log L(\theta)}{\partial \beta_1} \end{bmatrix}.$$

An arbitrary element of this score vector (i.e. with respect to either β_0 or β_1) can be expressed more generally like this: for $j = 0, 1$,

$$\begin{aligned}\frac{\partial \log L(\theta)}{\partial \beta_j} &= \sum_{i=1}^{n} \left[y_i \frac{\partial \log \theta_i}{\partial \theta_i} + (1 - y_i) \frac{\partial \log(1 - \theta_i)}{\partial \theta_i} \right] \\ &= \sum_{i=1}^{n} \left[y_i \frac{1}{\theta_i} \frac{\partial \theta_i}{\partial \beta_j} - \frac{1 - y_i}{1 - \theta_i} \frac{\partial \theta_i}{\partial \beta_j} \right] \\ &= \sum_{i=1}^{n} \left[y_i \frac{1}{\theta_i} - \frac{1 - y_i}{1 - \theta_i} \right] \frac{\partial \theta_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \left(\frac{y_i - \theta_i}{\theta_i (1 - \theta_i)} \right) \frac{\partial \theta_i}{\partial \beta_j},\end{aligned}$$

where the expression in brackets in the final line comes from replicating the working used to obtain equation (1): the only difference here is that θ is replaced with θ_i. Otherwise, the arithmetic used to get there is the same.

In fact, the term in the brackets, when combined with the partial derivative outside the brackets, can be used to construct a *generalized residual* of sorts, but that kind of question will be left for a later question.

(iv.) i. In this case we have

$$\begin{aligned}\frac{\partial \theta_i}{\partial \beta_0} &= \frac{\partial \Phi(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{i,2})}{\partial (\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})} \frac{\partial (\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})}{\partial \beta_2} \\ &= \phi(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}) \times 1 \\ &= \phi_i x_{i,2},\end{aligned}$$

where this time $\phi_i = \phi(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})$. In a similar fashion before we can write

$$\frac{\partial \theta_i}{\partial \beta_0} = \phi_i, \qquad \frac{\partial \theta_i}{\partial \beta_1} = \phi_i x_{i,1}.$$

ii. In the general case, we can write the linear regression for an arbitrary number of regressors k as

$$\begin{aligned}y_i &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{ik} + u_i \\ &= \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{ik} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \\ &= x_i' \beta + u_i,\end{aligned}$$

where $x_{i1} = 1$ is the regressor associated with the intercept term. This will justify the notation that we use later.

Based on our answers from before, we can infer that the expressions for the partials that we got before can now be generalized: namely, for any particular coefficient β_j for $j = 1, 2, \ldots, k$, we can now write

$$\begin{aligned}\frac{\partial \theta_i}{\partial \beta_j} &= \phi(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{ik}) x_{ij} \\ &= \phi(x_i'\beta) x_{ij} \\ &= \phi_i x_{ij}.\end{aligned}$$

iii. Stacking everything together, the gradient vector of θ_i would look something like this:

$$\frac{\partial \theta_i}{\partial \beta} = \begin{bmatrix} \frac{\partial \theta_i}{\partial \beta_1} \\ \frac{\partial \theta_i}{\partial \beta_2} \\ \vdots \\ \frac{\partial \theta_i}{\partial \beta_k} \end{bmatrix} = \begin{bmatrix} \phi_i x_{i1} \\ \phi_i x_{i2} \\ \vdots \\ \phi_i x_{ik} \end{bmatrix} = \phi_i \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{bmatrix} = \phi(x_i'\beta) x_i = \phi_i x_i.$$

(f) (i.) In general we are modelling

$$\mathbf{P}(y_i = 1 | x_i) = F(x_i'\beta),$$

where $F(\cdot)$ is the CDF of the disturbance term u_i (symmetric around zero), with associated PDF $f(\cdot)$. To simplify things, just consider the classic model with an intercept and extra regressor, which has an implied conditional probability

$$\mathbf{P}(y_i = 1 | x_i) = F(\beta_0 + \beta_1 x_i).$$

The marginal effect would be to evaluate the effect on $\mathbf{P}(y_i = 1|x_i)$ from a small change in x_i. This is done by taking the partial derivative with respect to x_i:

$$\frac{\partial F(\beta_0 + \beta_1 x_i)}{\partial x_i} = f(\beta_0 + \beta_1 x_i)\beta_1.$$

Notice that unlike the linear probability model (which has a constant interpretation of the marginal effects), the value of the PDF $f(\cdot)$ changes depending on the value of x_i, so the marginal effect is not constant for these types of models.

(ii.) For relative marginal effects, let's consider the general regression model $y_i = x_i'\beta + u_i$. The relative marginal effect is the *ratio* of marginal effects between two regressors x_{ij} and x_{ik}, which can be written as follows:

$$\frac{\left(\frac{\partial F(x_i'\beta)}{\partial x_{ij}}\right)}{\left(\frac{\partial F(x_i'\beta)}{\partial x_{ik}}\right)} = \frac{f(x_i'\beta)\beta_j}{f(x_i'\beta)\beta_k} = \frac{\beta_j}{\beta_k},$$

which is always constant no matter what the values of the regressors are.

(iii.) From part (i) above, observe that by construction, the PDF $f(\cdot) \geq 0$. Since it's multiplied with the coefficient β_1 (or any β_j in general, really), the marginal effect will share the same sign as the coefficient associated with x_{ij} provided that $f(\cdot) > 0$, which is almost always the case.

(g) Three ways in which you can interpret the output from a probit regression model proceed as follows:

- **Method 1: Signs.** The most crude method of all: interpreting only the sign on the coefficient to evaluate the direction in which the marginal effect goes. The justification as to why we can do this is in part (e) above. This tells us pretty much nothing about the actual value of the marginal effect, but is good for an inspection if we happen to be checking our intuition when doing first passes on the data for, say, research, coursework assignments or projects to name a few.
- **Method 2: Evaluation at the mean.** Since the value of $f(x_i'\beta)$ varies depending on what x_i happens to be, one way to obtain an estimated marginal effect is to evaluate the PDF at the means of the regressors x_i. In the general case this is written as

$$\text{(Mean Partial Effect)}_j = \phi(\bar{x}'\beta)\beta_j.$$

  This does come with a clear caveat: there may be no such observations in the dataset that actually looks like the average (so the usual things like an average household having say 2.7 people, 1.21 bedrooms and 0.3 pet canaries). In this kind of case, the practical uses of such an interpretation are quite limited.
- **Method 3: Average marginal effects.** This is a commonly reported way to interpret the output from a probit regression model. The idea here is that we can let *every* observation have a say in the computation of the marginal effect by doing the following:
  - For each observation in the sample (i.e. $i = 1, 2, \ldots, n$), calculate $x_i'\beta$ and hence $\phi(x_i'\beta)$.
  - Average them.
  - Calculate the *average marginal effect*

$$\text{(Average Marginal Effect)}_j = \left[\frac{1}{n}\sum_{i=1}^{n}\phi(x_i'\beta)\right] \times \beta_j.$$

    From the way it is written, one can observe that this effectively takes the formula for the marginal effect, and replaces the PDF $\phi(x_i'\beta)$ with a *weighted average* that accepts contributions from every observation in the dataset.

Question 2: GLM Formalities

(a) Suppose in a sample of n observations, we have that $y_i = 1$ occurs n_1 times in the dataset. This means that for these n_1 observations, $y_i^* = x_i'\beta > 0$. The associated probability of

this event occurring is

$$\mathbf{P}(y_i = 1|x_i) = F(x_i'\beta).$$

We also have that $y_i = 0$ occurs n_2 times, so that $n_1 + n_2 = n$. Here, $y_i^* > 0$ so that we have probability $\mathbf{P}(y_i = 0|x_i) = 1 - F(x_i'\beta)$. The idea here is that the *joint* density of observing n_1 observations of $y_i = 0$ and n_2 observations of $y_i = 1$ can be written as

$$g(y|X) = \prod_{i=1}^{n} F(x_i'\beta)^{y_i}(1 - F(x_i'\beta))^{1-y_i}.$$

In this case, $g(y|X)$ is the joint density of obtaining the dataset $y = y_1, y_2, \ldots, y_n$ given the regressors X. To get this joint density, each individual y_i is independent, so the joint density is simply given by

$$g(y|X) = \mathbf{P}(y_i = 1 \text{ occurs } n_1 \text{ times}) \times \mathbf{P}(y_i = 0 \text{ occurs } n_2 \text{ times}).$$

Expressions for each of these terms can be obtained by just rearranging the dataset so that the first n_1 observations have $y_i = 1$ and the remainder have $y_i = 0$. Then multiplying everything together we have

$$\begin{aligned}
g(y|X) &= \underbrace{F(x_1'\beta)F(x_2'\beta)\ldots F(x_{n_1}'\beta)}_{n_1 \text{ times}} \times \underbrace{[1 - F(x_{n_1+1}'\beta)][1 - F(x_{n_1+2}'\beta)]\ldots[1 - F(x_n'\beta)]}_{n_2 \text{ times}} \\
&= \left(\prod_{i=1}^{n_1} F(x_i'\beta)^{y_i}\right)\left(\prod_{i=n_1+1}^{n} (1 - F(x_i'\beta))^{1-y_i}\right) \\
&= \left(\prod_{i=1}^{n} F(x_i'\beta)^{y_i}\right)\left(\prod_{i=1}^{n}(1 - F(x_i'\beta))^{1-y_i}\right) \\
&= \prod_{i=1}^{n} F(x_i'\beta)^{y_i}(1 - F(x_i'\beta))^{1-y_i}.
\end{aligned}$$

Getting from the second to third lines can be a bit tricky. To see why we can do this, observe that we can in fact write

$$F(x_i'\beta)^{y_i} = \begin{cases} F(x_i'\beta) & \text{if } y_i = 1 \\ 1 & \text{if } y_i = 0 \end{cases}$$

so that

$$\prod_{i=1}^{n_1} F(x_i'\beta)^{y_i} = \underbrace{F(x_1'\beta)F(x_2'\beta)\ldots F(x_{n_1}'\beta)}_{n_1 \text{ times}} \times \underbrace{1 \times 1 \times \cdots \times 1}_{n_2 \text{ times}} = \prod_{i=1}^{n} F(x_i'\beta)^{y_i}.$$

Similarly, we also have that

$$(1 - F(x_i'\beta))^{1-y_i} = \begin{cases} 1 & \text{if } y_i = 1 \\ 1 - F(x_i'\beta) & \text{if } y_i = 0 \end{cases}$$

so that with a very similar (i.e. almost identical) argument, you can also show that

$$\prod_{i=n_1+1}^{n}(1-F(x_i'\beta))^{1-y_i} = \prod_{i=1}^{n}(1-F(x_i'\beta))^{1-y_i}$$

which finishes justifying the lines of working that we have above.

(b) The working here is actually more or less identical to question 1, part (c) for the Bernoulli function. In this case, we replace θ with $F(x_i'\beta)$. But to save ourselves time with notation, denote

$$F_i \equiv F(x_i'\beta) \qquad \text{and} \qquad f_i \equiv f(x_i'\beta)$$

as we're going to be using these expressions a lot. For the sake of completeness, let's derive the log-likelihood anyway:

$$\begin{aligned}
\log L(\beta; y, X) &= \log g(y|X) \\
&= \log \prod_{i=1}^{n} F(x_i'\beta)^{y_i}(1-F(x_i'\beta))^{1-y_i} \\
&= \sum_{i=1}^{n} \log[F(x_i'\beta)^{y_i}(1-F(x_i'\beta))^{1-y_i}] \\
&= \sum_{i=1}^{n}[y_i \log F_i + (1-y_i)\log(1-F_i)],
\end{aligned}$$

using our definition of F_i above.

(c) For the score, we can write

$$\frac{\partial F_i}{\partial \beta} = f(x_i'\beta)x_i = f_i x_i.$$

The motivation for this expression comes from Question 1, part (d)-(iv). Then, we can write the score as

$$\begin{aligned}
S(\beta) = \frac{\partial \log L(\beta)}{\partial \beta} &= \sum_{i=1}^{n}\left[y_i \frac{\partial \log F_i}{\partial \beta} + (1-y_i)\frac{\partial \log(1-F_i)}{\partial \beta}\right] \\
&= \sum_{i=1}^{n}\left[y_i \frac{\partial \log F_i}{\partial F_i}\frac{\partial F_i}{\partial \beta} + (1-y_i)\frac{\partial \log(1-F_i)}{\partial F_i}\frac{\partial F_i}{\partial \beta}\right] \\
&= \sum_{i=1}^{n}\left[\frac{y_i}{F_i} f_i x_i - (1-y_i)\frac{1}{1-F_i} f_i x_i\right] \\
&= \sum_{i=1}^{n}\left[\frac{y_i}{F_i} - \frac{(1-y_i)}{1-F_i}\right] f_i x_i \\
&= \sum_{i=1}^{n}\left[\frac{y_i - F_i}{F_i(1-F_i)}\right] f_i x_i \\
&= \sum_{i=1}^{n}\left[\frac{(y_i - F_i)f_i}{F_i(1-F_i)}\right] x_i \\
&= \sum_{i=1}^{n} x_i \nu_i(\beta) \qquad \text{as required,}
\end{aligned}$$

where

$$\nu_i(\beta) = \frac{(y_i - F_i)f_i}{F_i(1-F_i)} = \frac{(y_i - F(x_i'\beta))f(x_i'\beta)}{F(x_i'\beta)(1-F(x_i'\beta))}$$

is called the *generalized residual* once we expand all the notation back in. This kind of expression should give you an idea as to why we might like to shorten the notation! As a note: this expression is in fact scalar, so it can freely move around within the summation.

(d) To calculate the expected score, let's first evaluate the expectation of $\nu_i \equiv \nu_i(\beta)$ conditional on x_i. Observe that F_i and f_i are both functions of x_i, so they can be moved outside the conditional expectation without too much issue:

$$\begin{aligned}
\mathbb{E}(\nu_i|x_i) &= \mathbb{E}\left[\frac{(y_i - F_i)f_i}{F_i(1-F_i)}\middle| x_i\right] \\
&= \frac{1}{F_i(1-F_i)}\mathbb{E}[(y_i - F_i)f_i|x_i] \\
&= \frac{f_i}{F_i(1-F_i)}\mathbb{E}[(y_i - F_i)|x_i] \\
&= \frac{f_i}{F_i(1-F_i)}\left(\mathbb{E}(y_i|x_i) - F_i\right) \\
&= \frac{f_i}{F_i(1-F_i)} \times 0 && \because \mathbb{E}(y_i|x_i) = F_i \\
&= 0.
\end{aligned}$$

Therefore, the expected score is

$$\begin{aligned}
\mathbb{E}[S(\beta)] &= \mathbb{E}\left[\sum_{i=1}^n x_i\nu_i\right] \\
&= \mathbb{E}\left[\mathbb{E}\left(\sum_{i=1}^n x_i\nu_i\middle| x_i\right)\right] && \text{(Law of Iterated Expectations)} \\
&= \mathbb{E}\left[\sum_{i=1}^n x_i\mathbb{E}(\nu_i|x_i)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n x_i \times 0\right] \\
&= 0.
\end{aligned}$$

Hence the score has zero expectation, as required.

(e) Part (i). To get an idea of what the conditional variance of the score is going to look like, observe that we can write it as

$$\mathrm{Var}(S(\beta)|X) = \mathbb{E}(S(\beta)S(\beta)'|X).$$

Since $S(\beta)$ is a summation, we'll need a property of summation:

$$\sum_{i=1}^n a_i \sum_{j=1}^n b_j = \sum_{i=1}^n \sum_{j=1}^n a_i b_j$$

so that we can write

$$\begin{aligned} S(\beta)S(\beta)' &= \sum_{i=1}^{n} x_i\nu_i \left(\sum_{j=1}^{n} x_j\nu_j \right)' \\ &= \sum_{i=1}^{n} x_i\nu_i \sum_{j=1}^{n} x_j'\nu_j \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j'\nu_i\nu_j. \end{aligned}$$

Note that in the second line, the transpose moves inside the sum using the property that $(A+B)' = A'+B'$. We're not done yet; while this double summation is reasonably useful, it's best to be cautious before taking the conditional expectation right now, because we could quite easily miss crucial details if we do.

Recall how a double sum is evaluated: you first evaluate the *inner sum* (in this case, the one that starts from $j = 1$ and work your way back to the outer sum (the $i = 1$ part). But notice the following:

- Suppose that we pick a specific value of i.
- Then, we'd evaluate the inner sum for every $j = 1, \ldots, n$.
- In the special case where $j = i$, we actually have a term that looks like this:

  $$x_i x_i'\nu_i^2 \qquad \text{when } j = i.$$

  In this case the conditional expectation of ν_i^2 need not be zero!
- For all other terms $j \neq i$, the summation at j would give back the term

  $$x_i x_j'\nu_i\nu_j,$$

  and taking the expectation of that conditional on all of our regressors X would likely give us something zero.

So basically, it would help if we separated our all the terms where $j = i$. Since there are n observations total, there are only n situations where this can happen. In this case, we can give these terms their own summation:

$$\sum_{i=1}^{n} x_i x_i'\nu_i^2.$$

As for whatever's left, we can just edit the bounds on the summation over j to just not include the i terms. This would give us

$$\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} x_i x_j'\nu_i\nu_j \qquad \text{or, for the less notationally pedantic,} \qquad \sum_{i=1}^{n}\sum_{j\neq i} x_i x_j'\nu_i\nu_j.$$

Adding these together gives back a slightly nastier, albeit convenient, form of the conditional variance of the score:

$$S(\beta)S(\beta)' = \sum_{i=1}^{n} x_i x_i' \nu_i^2 + \sum_{i=1}^{n} \sum_{j \neq i} x_i x_j' \nu_i \nu_j.$$

Part (ii). Now it remains to check what the conditional expectation of ν_i^2 is. To do this, observe that

$$\nu_i^2 = \left(\frac{(y_i - F_i) f_i}{F_i(1 - F_i)} \right)^2 = \frac{(y_i - F_i)^2 f_i^2}{F_i^2 (1 - F_i)^2},$$

where for convenience, observe that f_i^2 and the entire denominator are functions of x_i. Taking the conditional expectation of this over x_i we get

$$\begin{aligned}
\mathbb{E}(\nu_i^2 | x_i) &= \mathbb{E}\left[\frac{(y_i - F_i)^2 f_i^2}{F_i^2 (1 - F_i)^2} \middle| x_i \right] \\
&= \frac{\mathbb{E}[(y_i - F_i)^2 | x_i] f_i^2}{F_i^2 (1 - F_i)^2} \\
&= \frac{\mathrm{Var}(y_i | x_i) f_i^2}{F_i^2 (1 - F_i)^2} \\
&= \frac{F_i(1 - F_i) f_i^2}{F_i^2 (1 - F_i)^2} && \because \mathrm{Var}(y_i | x_i) = F_i(1 - F_i) \\
&= \frac{f_i^2}{F_i(1 - F_i)} \equiv \delta_i.
\end{aligned}$$

If you're finding it a bit tricky to get the intuition behind the fact that $\mathrm{Var}(y_i|x_i) = F_i(1 - F_i)$, it's because of the properties of the Bernoulli distribution. Namely: when conditioned on x_i, $F(x_i'\beta)$ behaves as a constant. But since this constant is a probability of an outcome that is either zero or one, properties of the Bernoulli distribution, such as the mean and variance, apply here – just in the conditional sense!

In any case this means that we can now complete the question:

$$\begin{aligned}
\mathrm{Var}(S(\beta)|X) &= \mathbb{E}(S(\beta)S(\beta)'|X) \\
&= \mathbb{E}\left(\sum_{i=1}^{n} x_i x_i' \nu_i^2 + \sum_{i=1}^{n} \sum_{j \neq i} x_i x_j' \nu_i \nu_j \middle| X \right) \\
&= \sum_{i=1}^{n} x_i x_i' \mathbb{E}(\nu_i^2 | x_i) + \sum_{i=1}^{n} \sum_{j \neq i} x_i x_j' \mathbb{E}(\nu_i | x_i) \mathbb{E}(\nu_j | x_j) \\
&= \sum_{i=1}^{n} x_i x_i' \mathbb{E}(\nu_i^2 | x_i) \\
&= \sum_{i=1}^{n} \delta_i x_i x_i' \\
&= X \Delta X',
\end{aligned}$$

where $\Delta = \text{diag}(\delta_1, \delta_2, \ldots, \delta_n)$.

Note: if you're having trouble figuring out why it can be written like this, just consider a two observation, two variable case. So x_i would be 2×1. Then the stacked data matrix is

$$X = \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}$$

so that one has, using the properties of partitioned matrix transposes,

$$X'\Delta X = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix} \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} \delta_1 x_1' \\ \delta_2 x_2' \end{bmatrix} = \delta_1 x_1 x_1' + \delta_2 x_2 x_2' = \sum_{i=1}^{n} \delta_i x_i x_i',$$

which will hopefully help with your general understanding.

(f) Part (i). Let γ_i and θ_i be as defined in the question. First, expand out the brackets:

$$\gamma_i = y_i f_i - F_i f_i \qquad \text{and} \qquad \theta_i = F_i - F_i^2.$$

For γ_i we will need to use the Product Rule for the expression $F_i f_i$, since it is a product of two functions of β. Furthermore, note that the Chain Rule also has to be used on partials taken with respect to f_i:

$$\frac{\partial f_i}{\partial \beta_j} = \frac{\partial f(x_i'\beta)}{\partial \beta_j} = \frac{\partial f(x_i'\beta)}{\partial x_i'\beta} \frac{\partial x_i'\beta}{\partial \beta_j} = \frac{\partial f_i}{\partial x_i'\beta} x_{ij}.$$

Therefore we have

$$\begin{aligned}
\Gamma_{ij} &= \frac{\partial \gamma_i}{\partial \beta_j} \\
&= y_i \frac{\partial f_i}{\partial \beta_j} - \frac{\partial}{\partial \beta_j} F_i f_i \\
&= y_i \frac{\partial f_i}{\partial x_i'\beta} x_{ij} - \underbrace{\left(F_i \frac{\partial f_i}{\partial \beta_j} + f_i \frac{\partial F_i}{\partial \beta_j} \right)}_{\text{Product Rule}} \\
&= y_i \frac{\partial f_i}{\partial x_i'\beta} x_{ij} - \left(F_i \frac{\partial f_i}{\partial x_i'\beta} x_{ij} - f_i^2 x_{ij} \right) \qquad \because \frac{\partial F_i}{\partial \beta_j} = f_i x_{ij} \\
&= y_i \frac{\partial f_i}{\partial x_i'\beta} x_{ij} - F_i \frac{\partial f_i}{\partial x_i'\beta} x_{ij} - f_i^2 x_{ij} \\
&= \left(y_i \frac{\partial f_i}{\partial x_i'\beta} - F_i \frac{\partial f_i}{\partial x_i'\beta} - f_i^2 \right) x_{ij} \\
&= \left([y_i - F_i] \frac{\partial f_i}{\partial x_i'\beta} \right) x_{ij},
\end{aligned}$$

as required. Now with one out of the way, it's time for the other one. For this one, note that we need to use the Chain Rule twice for one of our terms: namely,

$$\begin{aligned}
\frac{\partial F_i^2}{\partial \beta_j} &= \frac{\partial F(x_i'\beta)^2}{\partial \beta_j} = \frac{\partial F(x_i'\beta)^2}{\partial F(x_i'\beta)} \frac{\partial F(x_i'\beta)}{\partial x_i'\beta} \frac{\partial x_i'\beta}{\partial \beta_j} \\
&= 2F(x_i'\beta) \times f(x_i'\beta) \times x_{ij} \\
&= 2F_i f_i x_{ij}
\end{aligned}$$

in our shorthand notation. From here we can get back to the main task:

$$\begin{aligned}\Theta_{ij} &= \frac{\partial \theta_i}{\partial \beta_j} \\ &= \frac{\partial F_i}{\partial \beta_j} - \frac{\partial F_i^2}{\partial \beta_j} \\ &= f_i x_{ij} - 2F_i f_i x_{ij} \\ &= (1 - 2F_i) f_i x_{ij},\end{aligned}$$

as required.

All that remains now is to handle the partial of ν_i with respect to β_j. But since that's a fraction, we're going to have to use the Quotient Rule, which is not going to be pretty:

$$\begin{aligned}\frac{\partial \nu_i}{\partial \beta_j} &= \frac{\theta_i \Gamma_{ij} - \gamma_i \Theta_{ij}}{\theta_i^2} \\ &= \frac{F_i(1-F_i)\left[(y_i - F_i)\dfrac{\partial f_i}{\partial x_i'\beta} - f_i^2\right] x_{ij} - (y_i - F_i) f_i (1 - 2F_i) f_i x_{ij}}{F_i^2 (1-F_i)^2} \\ &= \underbrace{\frac{F_i(1-F_i)\left[(y_i - F_i)\dfrac{\partial f_i}{\partial x_i'\beta} - f_i^2\right] - (y_i - F_i)(1-2F_i) f_i^2}{F_i^2(1-F_i)^2}}_{\equiv \alpha_i} x_{ij} \\ &= \alpha_i x_{ij}.\end{aligned}$$

Part (ii). To figure out what we can use this partial expression for, observe that the Hessian is defined by

$$H(\beta) = \frac{\partial}{\partial \beta'} S(\beta).$$

The key part to note is the β' in the partial notation. I usually find that a lot of people have plenty of trouble figuring out exactly why the transpose is there, so the aside below will talk about this.

Aside on vector calculus dimensions. One thing that you're probably used to already is the idea of the gradient vector:

$$S(\beta) = \frac{\partial \log L(\beta)}{\partial \beta}$$

where $\log L(\beta)$ is a *scalar* term and β is a $k \times 1$ column vector of parameters. The end result is a gradient vector that is also $k \times 1$. So the key point here is noting the following:

*The gradient of a **scalar** with respect to a **column vector***
*gives a **column vector**.*

Similarly, if we took the gradient with respect to a *row vector* it would similarly be a row vector. If you're happy with this, then this gives us the tools that we need to conceptualize the Hessian:

- Consider some arbitrary element of the gradient vector.
- Since it's a single element, it's naturally a scalar!
- Now, take the gradient of that with respect to β'.
- Since β' is a row vector (i.e. the transpose of a column vector), the gradient of this element will also be a row vector.
- Repeat this for every single element in the gradient vector.

The Hessian is obtained by **stacking** all of these individual gradient vectors, which are now all in row format.[1]

Back to it. So in any case, what does the Hessian give us? First, let's examine what happens if we take the gradient of ν_i with respect to β' for a simplified case of, say, two elements in β. In this case we'd get something looking like this:

$$\begin{aligned}\frac{\partial \nu_i}{\partial \beta'} = \begin{bmatrix}\frac{\partial \nu_i}{\partial \beta_1} & \frac{\partial \nu_i}{\partial \beta_2}\end{bmatrix} &= \begin{bmatrix}\alpha_i x_{i1} & \alpha_i x_{i2}\end{bmatrix} \\ &= \alpha_i \begin{bmatrix} x_{i1} & x_{i2}\end{bmatrix} \\ &= \alpha_i x_i'.\end{aligned}$$

Now we can extend this to the more general case of k coefficients. The Hessian is therefore

$$\begin{aligned}H(\beta) = \frac{\partial S(\beta)}{\partial \beta'} &= \frac{\partial}{\partial \beta'}\left[\sum_{i=1}^{n} x_i \nu_i\right] \\ &= \left[\sum_{i=1}^{n} x_i \frac{\partial \nu_i}{\partial \beta'}\right] \\ &= \sum_{i=1}^{n} x_i \alpha_i x_i' \\ &= \sum_{i=1}^{n} \alpha_i x_i x_i',\end{aligned}$$

since α_i is a scalar.

(g) If we take the conditional expectation of $H(\beta)$ over the regressors x_i, we'd get this expression:

$$\mathbb{E}(H(\beta)|X) = \mathbb{E}\left[\sum_{i=1}^{n} \alpha_i x_i x_i'\right] = \sum_{i=1}^{n} \mathbb{E}(\alpha_i|x_i) x_i x_i'.$$

[1] Note that this also implies another rule of vector calculus: the gradient of a $k \times 1$ vector with respect to a $1 \times m$ vector gives back a $k \times m$ matrix. But since we don't really need to pay too much attention to that it gets to be here in the footnote instead.

So our answer will revolve around precisely what the expected value of α_i given x_i is. Unfortunately, this means that we've got some tedious work ahead of us. Noting that F_i and f_i, along with their squared versions, are all functions of x_i, all we need to do is focus on y_i. Doing this gives us:

$$\begin{aligned}
\mathbb{E}(\alpha_i|x_i) &= \mathbb{E}\left[\left.\frac{F_i(1-F_i)\left[(y_i-F_i)\dfrac{\partial f_i}{\partial x_i'\beta} - f_i^2\right] - (y_i-F_i)(1-2F_i)f_i^2}{F_i^2(1-F_i)^2}\right| x_i\right] \\
&= \frac{1}{F_i^2(1-F_i)^2}\left(F_i(1-F_i)\left[(\mathbb{E}(y_i|x_i)-F_i)\frac{\partial f_i}{\partial x_i'\beta} - f_i^2\right] - (\mathbb{E}(y_i|x_i)-F_i)(1-2F_i)f_i^2\right) \\
&= \frac{1}{F_i^2(1-F_i)^2}\left(F_i(1-F_i)[0-f_i^2] - 0\times(1-2F_i)f_i^2\right) \\
&= \frac{1}{F_i^2(1-F_i)^2}\left(F_i(1-F_i)(-f_i^2)\right) \\
&= -\frac{f_i^2}{F_i^2(1-F_i)^2} \\
&= -\delta_i.
\end{aligned}$$

Note that in the third line, we invoke the property that $\mathbb{E}(y_i|x_i) = F_i$, which results in their difference being zero. Therefore, we have

$$\mathbb{E}(H(\beta)|X) = \sum_{i=1}^{n}\mathbb{E}(\alpha_i|x_i)x_ix_i' = -\sum_{i=1}^{n}\delta_i x_i x_i' = -\text{Var}(S(\beta)|X).$$

This turns out to be the usual information matrix equality, but in the conditional sense instead. Quite a lot of work to verify something that we've been working with for quite a while, but hopefully this will help your understanding of the underlying content behind binary response models!