4 Likelihood-Based Inference

The previous chapter dealt with OLS estimation of the linear regression model, i.e. the conditional expectation $\mathbb{E}[y_i|x_i, \theta]$ is specified to be linear in the parameters θ, and – out of all properties of the distribution of data $p(y_i, x_i|\theta)$ – the estimation method only requires specifying this conditional expectation in order to obtain the estimator $\hat{\beta}_{OLS}$.[1] This chapter studies likelihood-based inference methods, which require the researcher to specify the whole (conditional) distribution $p(y_i|x_i, \theta)$. With the cost of more assumptions than under OLS, these methods can conduct inference for a wider class of models and allow the researcher to use additional hypothesis tests. In addition, specifying the whole distribution of data enables Bayesian inference.

This chapter is set out as follows. First, Section 4.1 discusses Maximum Likelihood (ML) estimation of the linear regression model. Then, Section 4.2 shows some properties of ML that apply for more generalm models as well, Section 4.3 presents likelihood-based hypothesis tests, and Section 4.4 mentions a few (cross-sectional) models commonly estimated using ML. In turn, Section 4.5 discusses Bayesian estimation of the linear regression model, which includes Lasso and Ridge estimation. Following up, Section 4.6 explains how Bayesian inference can be implemented for more general models (likelihood and prior specifications), and Section 4.7 discusses model selection. As before, unless otherwise stated, the frequentist paradigm applies, i.e. we treat θ as a fixed parameter and condition all moments and distributions on it, though this conditioning is often omitted for notational simplicity.

[1] To obtain its variance and asymptotic distribution, also $\mathbb{V}[y_i|x_i, \theta] = \mathbb{V}[u_i|x_i, \theta] = \sigma^2$ had to be specified. Also, as mentioned in the previous chapter, the assumption $\mathbb{E}[x_i u_i|\theta] = \mathbb{E}[x_i(y_i - x_i'\beta)|\theta] = 0$ – less restrictive than $\mathbb{E}[y_i|x_i, \theta] = x_i'\beta$ – preserves most properties of the OLS estimator. Thereby, $\theta = \{\beta, \sigma^2\}$.

4.1 Maximum Likelihood Inference for the Linear Regression Model

As in the previous chapter, suppose a scalar y_i is linearly related to a vector of regressors x_i. Suppose in addition that the error term follows a Normal distribution:

$$y_i = x_i'\beta + u_i \;, \quad u_i|x_i \overset{i.i.d.}{\sim} N(0, \sigma^2) \;.$$

This implies that $y_i|x_i, \theta \overset{i.i.d.}{\sim} N(x_i'\beta, \sigma^2)$ with $\theta = (\beta', \sigma^2)'$,[2] i.e.

$$p(y_i|x_i, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2 \right\} \;.$$

The likelihood function is defined as the joint pdf of all observations, which by independence is equal to the product of the marginal pdfs of the individual observations:

$$\mathcal{L}(\theta|Y, X) \equiv p(Y|X, \theta) = \prod_{i=1}^{n} p(y_i|x_i, \theta) \;.$$

Algebraic manipulations allow us to write the likelihood out explicitly:

$$\begin{aligned} \mathcal{L}(\theta|Y, X) \equiv p(Y|X, \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{ -\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2 \right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} exp\left\{ -\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) \right\} \;. \end{aligned}$$

The log-likelihood is then

$$\begin{aligned} \ell(\theta|Y, X) &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) \\ &= c - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) \;, \end{aligned}$$

where c is a constant that does not depend on θ.

The ML estimator for θ, $\hat{\theta}_{ML}$, is defined as the value for θ that maximizes $\ell(\theta|Y, X)$:

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} \ell(\theta|Y, X) \;.^3$$

[2] Note that independent sampling across i and homoskedasticity are subsumed in $u_i|x_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

For the linear regression model, this leads to the FOCs

$$[\beta]: \quad -\frac{1}{2\sigma^2}2X'(Y-X\beta) = 0 \ , \qquad [\sigma^2]: \quad \frac{n}{2}\frac{1}{\sigma^2} + \frac{2(Y-X\beta)'(Y-X\beta)}{(2\sigma^2)^2} = 0 \ ,$$

which imply

$$\hat{\beta}_{ML} = (X'X)^{-1}X'Y \ , \quad \hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_{ML})^2 \ .$$

Note that $\hat{\beta}_{ML} = \hat{\beta}_{OLS}$. As a result, $\hat{\beta}_{ML}$ inherits all the properties of $\hat{\beta}_{OLS}$. In particular, it has the same, Normal asymptotic distribution. In addition, however, under the assumption $u_i|x_i$ is Normal, we get that $\hat{\beta}_{ML}$ (conditional on X)[4] follows a Normal distribution already in finite samples:

$$\hat{\beta}_{ML}|X \sim N\left(\beta \ , \ \frac{\sigma^2}{n}\hat{Q}^{-1}\right) \ , \quad \text{with } \hat{Q} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \ .$$

To see this, note that if the model is specified correctly, i.e. $y_i = x_i'\beta + u_i$, then $\hat{\beta}_{ML}$ is a weighted sum of Normal RVs: $\hat{\beta}_{ML} = \beta + \hat{Q}^{-1}\frac{1}{n}\sum_{i=1}^{n} x_i u_i$. Thereby, $\mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} x_i u_i|X] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[x_i u_i|x_i] = \frac{1}{n^2}\sum_{i=1}^{n} x_i x_i'\mathbb{E}[u_i^2|x_i] = \frac{\sigma^2}{n}\hat{Q}$.[5] As a result, if $u_i|x_i$ is Normal, the t-test-statistic $T_t(x)$ not only converges in distribution to $N(0,1)$, but it exactly follows a $N(0,1)$ for any sample size n.

4.2 Maximum Likelihood - Finite Sample Results

This section establishes some useful finite sample properties of the ML estimator

$$\hat{\theta}_{ML} = \arg\max_{\theta} \ell(\theta|Y)$$

for general models (i.e. general likelihood functions). Asymptotic properties are discussed as a particular example of the more general extremum estimation theory in Chapter 5. For notational simplicity, X is dropped from the conditioning set in $\ell(\theta|Y) = \log p(Y|\theta)$.[6]

[3] Either we define the likelihood as $\mathcal{L}(\theta|Y,X) = p(Y \ |X,\theta)$, or we let $\mathcal{L}(\theta|Y,X) = p(Y,X|\theta) = p(Y|X,\theta)p(X|\theta)$ and assume that the marginal density of X does not depend on θ, which makes it irrelevant for our inference problem and allows us to maximize only $p(Y|X,\theta)$. The MLE approach is sometimes motivated based on the Kullback-Leibler inequality rather than heuristically as the maximization of the probability of obtaining the realizations in one's sample (see Appendix).

[4] As well as σ^2, but this is implied here as under frequentist approaches such as ML we treat $\theta = (\beta', \sigma^2)'$ as a fixed parameter.

[5] We can also obtain the distribution of $\hat{\beta}_{ML}|X$ using an approach relying on the concept of a sufficient statistic (see Appendix).

[6] That is, Y is all the data that is considered random.

Define the score of the likelihood function as the $k \times 1$-vector of first-order derivatives of the log-likelihood function:

$$s(\theta) \equiv \ell'(\theta|Y) = \frac{\partial}{\partial\theta}\ell(\theta|Y) \; .$$

For example, in the linear regression model if σ^2 is known and we are only interested in estimating β, we have $s(\beta) = \frac{1}{\sigma^2}X'(Y - X\beta)$. The Hessian of the likelihood function is the $k \times k$-matrix of second-order derivatives of the log-likelihood function:

$$H(\theta) = \ell^{''}(\theta|Y) = \frac{\partial^2}{\partial\theta'\theta'}\ell(\theta|Y) \; .$$

For the linear regression model, $H(\beta) = -\frac{1}{\sigma^2}X'X$.

Proposition 22 (Information Matrix Equality).
$I(\theta_0) = \mathbb{E}[s(\theta_0)s(\theta_0)'] = -\mathbb{E}[H(\theta_0)]$, where $I(\theta_0)$ is called the information matrix.[7]

Note that it holds for the linear regression model: $\mathbb{E}[s(\beta_0)s(\beta_0)'] = \mathbb{E}[\frac{1}{\sigma^2}X'UU'X\frac{1}{\sigma^2}] = \frac{1}{\sigma^2}X'X = -\mathbb{E}[H(\beta_0)]$. Evaluating these expressions at the supposedly true parameter value, θ_0, is important because $\mathbb{E}[s(\tilde{\beta})s(\tilde{\beta})'] = \mathbb{E}[\frac{1}{\sigma^2}X'(Y - X\tilde{\beta})(Y - X\tilde{\beta})'X\frac{1}{\sigma^2}] \neq \mathbb{E}[\frac{1}{\sigma^2}X'UU'X\frac{1}{\sigma^2}] = \frac{1}{\sigma^2}X'X = -\mathbb{E}[H(\tilde{\beta})]$ for any $\tilde{\beta} \neq \beta_0$.

Proposition 23 (Cramer-Rao Lower Bound).
Let $\tilde{\theta}$ be an unbiased estimator for θ_0. Then $V[\tilde{\theta}] \geqslant I^{-1}(\theta_0)$.[8]

In other words, the inverse of the information matrix provides the so-called Cramer-Rao lower bound for the variance of unbiased estimators. For intuition, consider the case when θ is a scalar. Then $I(\theta_0)$ is a scalar too. It is the second-order derivative of the likelihood function w.r.t. θ, evaluated at θ_0. If on average, averaging over all possible observations, it gives a high number, then the likelihood function is very peaked, which means that the likelihood contains a lot of information on where the true θ_0 lies.

The ML estimator often attains the Cramer-Rao lower bound, hence being the most efficient estimator (i.e. the estimator with the lowest variance) among all unbiased estimators. For example, note that this holds true for the linear regression model, where $\mathbb{V}[\hat{\theta}_{ML}] = \mathbb{V}[\hat{\theta}_{OLS}] = I^{-1}(\theta_0) = \sigma^2(X'X)^{-1}$. This result goes farther than the Gauss-Markov theorem, as it does not require linearity of the estimator, nor does it require a linear model to begin with.

[7] While it holds under more general conditions, the Appendix contains a proof for the case where θ is a scalar and $p(Y|\theta)$ is sufficiently smooth s.t. we can exchange the order of integration and differentiation.

[8] The appendix contains a heuristic proof for the case where θ_0 is a scalar, but this result too holds for rather general conditions.

Note that we have not only $\mathbb{E}[s(\theta_0)s(\theta_0)'] = -\mathbb{E}[H(\theta_0)]$ for $s(\theta) = \ell'(\theta|Y) = \frac{\partial p(Y|\theta)}{\partial \theta}$ and $H(\theta) = \ell''(\theta|Y) = \frac{\partial^2 p(Y|\theta)}{\partial\theta\partial\theta'}$, but also

$$\mathbb{E}[s_i(\theta_0)s_i(\theta_0)'] = -\mathbb{E}[H_i(\theta_0)]$$

for $s_i(\theta) = \frac{\partial p(y_i|\theta)}{\partial \theta}$ and $H(\theta) = \frac{\partial^2 p(y_i|\theta)}{\partial\theta\partial\theta'}$. For example, in the linear regression model, $s_i(\beta) = \frac{1}{\sigma^2}x_i(y_i - x_i'\beta)$ and $H_i(\beta) = -\frac{1}{\sigma^2}x_i x_i'$, and it holds that $\mathbb{E}[s_i(\beta_0)s_i(\beta_0)'] = \mathbb{E}[\frac{1}{\sigma^4}x_i u_i(x_i u_i)'] = \mathbb{E}[\frac{1}{\sigma^2}x_i x_i'] = -\mathbb{E}[H_i(\beta_0)]$. This result is useful for assessing the asymptotic properties of ML estimators (see Chapter 5).

Quasi Maximum Likelihood Estimation If the distribution of $Y|X$ is actually as supposed in the derivation of the likelihood function (often Normal, as in the analysis above), we call $\mathscr{L}(\theta)$ the likelihood function and $\hat{\theta} = \arg\max_\theta \mathscr{L}(\theta)$ the ML estimator. If we do not really believe in this Normality assumption, but use it anyway to conduct an ML analysis, then these objects are called the quasi likelihood function and Quasi-ML estimator (QMLE). Such a QMLE analysis can be conducted for several reasons. First, some models can only be estimated using ML, and the ML estimator is well-behaved even if the supposed distribution is misspecified. For example, in the linear regression model, even if $u_i|x_i$ is not exactly Normal, we know that $\hat{\beta}_{ML} = \hat{\beta}_{OLS}$ has still good properties. Second, specifying the likelihood allows one to use likelihood-based hypothesis tests (and confidence set constructions) and, third, it enables Bayesian inference.

4.3 Likelihood-Based Testing

Section 2.3 introduced the hypothesis testing problem and discussed as examples the t-test and Likelihood Ratio (LR) test in a simple setting. Section 3.3 revisited the t-test in the context of the linear regression model and introduced the Wald test. Likelihood-based inference enables further testing approaches.

Suppose, as before, that we want to test $\mathscr{H}_0 : g(\theta) = 0$ vs. $\mathscr{H}_1 : g(\theta) \neq 0$ for some function $g : \mathbb{R}^k \to \mathbb{R}^m$ (i.e. $m \leq k$ restrictions). Let

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \ell(\theta|Y) \ , \quad \text{and} \quad \bar{\theta}_n = \arg\max_{\theta\in\Theta, g(\theta)=0} \ell(\theta|Y)$$

be the unrestricted and restricted ML estimators, respectively.

As discussed in the derivation of the Wald test in the Appendix of Chapter 3, it can be applied as long as g has continuous first derivatives and $\hat{\theta}$ is asymptotically Normal, i.e. $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$ for some V. Therefore, it also applies to the ML estimator for a wide class of models (see Chapter 5 for asymptotic Normality of the ML estimator). Recall that it is based on the statistic

$$T_W = n\, g(\hat{\theta})' \left[G(\hat{\theta}) \hat{V} G(\hat{\theta})' \right]^{-1} g(\hat{\theta}) \xrightarrow{d} \chi^2_m \;,$$

where $\hat{V}$ is a consistent estimator for V and $G(\hat{\theta}) = \partial g(\theta)/\partial\theta \mid_{\theta=\hat{\theta}}$ is the $m \times k$ matrix of derivatives of g with respect to θ evaluated at $\hat{\theta}$. Since under $\mathscr{H}_0$, both $\hat{\theta}$ and $\bar{\theta}$ are valid estimators of θ_0, either can be used to construct the test statistic above.

Relative to OLS, ML estimation enables two more testing approaches, which exploit different implicatons of $\mathscr{H}_0$ being true. First, as introduced in Section 2.3, the likelihood ratio (LR) test uses the statistic

$$T_{LR} = 2 \left[\ell(\hat{\theta}|Y) - \ell(\bar{\theta}|Y) \right] \xrightarrow{d} \chi^2_m \;.$$

Second, the Lagrange Multiplier (LM) test uses the statistic

$$T_{LM} = s(\bar{\theta})' I(\bar{\theta})^{-1} s(\bar{\theta}) \xrightarrow{d} \chi^2_m \;.^{9}$$

The Wald, LR and LM tests are based on three different results that should hold true under $\mathscr{H}_0$: i) the difference between $g(\hat{\theta})$ and $g(\theta) = 0$ should be small, ii) the difference between the likelihood evaluated at the restricted and unrestricted ML estimator should be small, and iii) the slope of the likelihood at the restricted ML estimator, $s(\bar{\theta})$, should be close to zero. Fig. 4.1 illustrates. By the Neyman-Pearson Lemma (Proposition 18) we know that if both $\mathscr{H}_0$ and $\mathscr{H}_1$ are point hypotheses, the LR test is optimal. Even for composite $\mathscr{H}_0$ and $\mathscr{H}_1$ it usually performs very well. However, it has the drawback that one needs to compute both the restricted and unrestricted ML estimator, which might be difficult for some models. In contrast, the LM test requires the researcher only to obtain the restricted ML estimator, while the Wald test only uses the unrestricted ML estimator. The Wald test has the disadvantage that for some $\mathscr{H}_0$, one can change the outcome of the test by changing the function g. For example, $g(\theta) = \theta - 1$ and $g(\theta) = log(\theta)$ both test $\theta = 1$ but can lead to different test outcomes.

As an example, suppose we test $\mathscr{H}_0 : \beta = \beta_0$ in the linear regression model $y_i = x_i'\beta + u_i$, $u_i|x_i \sim N(0,1)$. We can use the function $g(\beta) = \beta - \beta_0$ with $g'(\beta) = I_k$. We know

[9] It is based on the fact that $\frac{1}{\sqrt{n}} s(\theta_0) \xrightarrow{d} N(0, \frac{1}{n} I(\theta_0))$. See Chapter 5.

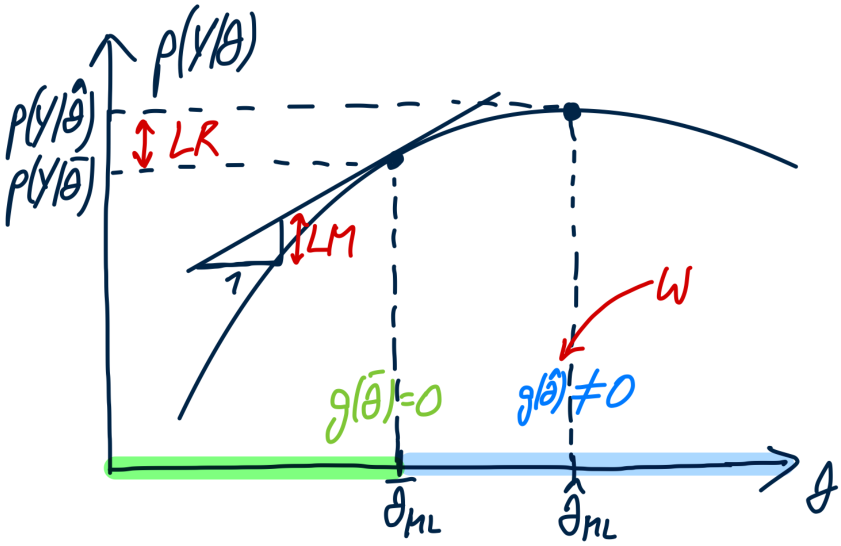

Figure 4.1: Wald, LR & LM Tests

Notes: Illustration of the Wald, LR & LM tests for a scalar parameter θ.

$\ell(\beta|Y,X) \propto -\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$ and $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0, Q^{-1})$ with $Q = \mathbb{E}[x_i x_i{}']$. We get

$$
\begin{aligned}
T_{LR} &= \sum_{i=1}^{n}(y_i - x_i'\beta_0)^2 - \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 \ , \\
T_W &= n(\hat{\beta}-\beta_0)'\hat{Q}(\hat{\beta}-\beta_0) \ , \\
T_{LM} &= \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(y_i - x_i'\beta_0)x_i\right]' \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta_0)^2 x_i x_i'\right]^{-1} \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(y_i - x_i'\beta_0)x_i\right] \ .
\end{aligned}
$$

One can show that $T_{LR} = T_W$ in this case.

4.4 Maximum Likelihood Applications

This section discusses several cross-sectional models which are typically estimated by ML. It focuses on the rationale behind these particular models and on how to obtain the likelihood function. Maximizing the latter, one obtains the ML estimator. When this is not possible analytically, the ML estimator is obtained using numerical optimization methods, which are outlined in Section 7.1. Asymptotic properties of the ML estimator are discussed in Chapter 5.

4.4.1 Binary Choice: Probit & Logit Model

Suppose $y_i \in \{0,1\}$, i.e. the outcome variable can only take the values zero or one. In this case, linear regression is not attractive because it does not take into account that $\mathbb{E}[y_i|x_i] = \mathbb{P}[y_i = 1|x_i]$ is bounded between zero and one, but can generate predicted outcomes that are below zero or exceed one. This is related to the fact that it implies constant partial effects, i.e. a given change in x_i always translates into the same change of y_i, no matter the values of x_i and y_i (i.e. no matter whether y_i is already close to one of the boundaries).

The common approach to dealing with this case is to set up a latent variable model:

$$y_i^* = x_i'\beta + u_i \;\;, \quad u_i|x_i \sim N(0,1) \;,$$

and assume we observe $y_i = \mathbf{1}\{y_i^* \geqslant 0\}$. For example, y_i^* could be the underlying utility that determines a consumer's buying choice $y_i \in \{0,1\}$. If this utility is positive, we observe $y_i = 1$, if it is negative, we observe $y_i = 0$.

For notational simplicity, I omit x_i (and β) from the conditioning set in the following calculations. We have

$$\mathbb{P}[y_i = 1] = \mathbb{P}[x_i'\beta + u_i \geqslant 0] = \mathbb{P}[u_i \geqslant -x_i'\beta] = 1 - \Phi(-x_i'\beta) = \Phi(x_i'\beta)$$

and $\mathbb{P}[y_i = 0] = 1 - \Phi(x_i'\beta) = \Phi(-x_i'\beta)$, where $\Phi(x)$ is the cdf of a standard Normal RV. Hence, y_i has the pdf

$$p(y_i|x_i,\beta) = \begin{cases} \Phi(x_i'\beta) & \text{if } y_i = 1 \\ \Phi(-x_i'\beta) & \text{if } y_i = 0 \end{cases} \quad = \Phi(x_i'\beta)^{y_i}\Phi(-x_i'\beta)^{1-y_i}$$

(it is Bernoulli-distributed with probability of success $\Phi(x_i'\beta)$). This leads to the likelihood

$$\mathcal{L}(\beta|Y,X) = \prod_{i=1}^{n} \Phi(x_i'\beta)^{y_i}\Phi(-x_i'\beta)^{1-y_i} \;,$$

and the log-likelihood $\ell(\beta|Y,X) = \sum_{i=1}^n y_i \log(\Phi(x_i'\beta)) + (1-y_i)\log(\Phi(-x_i'\beta))$.[10]

Maximizing this likelihood yields $\hat{\beta}_{ML}$. As no analytical solution is available, this maximization is conducted numerically. Given $\hat{\beta}_{ML}$, we can compute fitted values $\hat{y}_i = \mathbb{E}[y_i|x_i,\hat{\beta}] = \Phi(x_i'\hat{\beta})$. The R^2 statistic can then be computed as $1 - SSR/SST$, where SSR is the sum of

[10] Note that we can fix the variance of $u_i|x_i$ to one without loss of generality. This is because if $u_i|x_i \sim N(0,\sigma^2)$, we would have $\frac{\beta}{\sigma}$ instead of β in the calculations above. This means that we can only identify $\frac{\beta}{\sigma}$, but not β and σ separately.

squared residuals $y_i - \hat{y}_i = y_i - \Phi(x_i'\hat{\beta})$.

Partial Effects If x_i goes from x_1 to x_2, its partial effect on y_i is

$$\begin{aligned}\mathbb{E}[y_i|x_i = x_2] - \mathbb{E}[y_i|x_i = x_1] &= \mathbb{P}[y_i = 1|x_i = x_2] - \mathbb{P}[y_i = 1|x_i = x_1] \\ &= \Phi(x_2'\beta) - \Phi(x_1'\beta) \ .\end{aligned}$$

For small changes, it can be approximated as $\frac{\partial \mathbb{E}[y_i|x_i=x]}{\partial x} = \phi(x'\beta)\beta$. Note that the partial effect depends on x; it is large for "central", moderate values of x and is smaller for more extreme values of x. To estimate these partial effects, we replace β with $\hat{\beta}$.

Since $\Phi(\cdot)$ is a strictly increasing function, the sign of β_j reveals the sign of the partial effect of x_j, but the size of β_j is not interpretable. Only the relative sizes of two coefficients β_k and β_l have (qualitative) meaning. Nevertheless, we can test for the partial effect of x_j being zero by testing $\mathscr{H}_0 : \beta_j = 0$, because the former is zero iff $\beta_j = 0$.

Probit vs Logit Different distributional assumptions for u_i lead to different binary choice models. If, as here, we assume that $u_i|x_i \sim N(0,1)$, we get the probit model. If we assume $u_i|x_i \sim L(0,1)$ (standard logistic distribution), we get the logit model (or logistic regression). Then the cdf of $u_i|x_i$ is $L(x) = 1/(1 + exp\{-x\})$ instead of $\Phi(x)$.

4.4.2 Censored Outcomes: Tobit Model

Suppose the observed outcome variable y_i is censored at zero: $y_i \geq 0$. We can deal with this case with the following latent variable model:

$$y_i^* = x_i'\beta + u_i \ , \quad u_i|x_i \sim N(0,\sigma^2) \ , \quad y_i = y_i^* \ \mathbf{1}\{y_i^* \geqslant 0\} \ .$$

For example, y_i^* could be the desired interest rate of a central bank which cannot set negative interest rates. More generally, we might have $y_i \geq \gamma$, and so we set $y_i^* \geqslant \gamma$. In some applications, γ is a parameter to be estimated. For example, say y_i is an individual's wage, and we only observe the wage of people who get offered a wage higher than the reservation wage γ.

The standard linear regression model is unattractive because it does not take into account this lower bound on y_i. However, note that if y_i was positive for all i – $y_i > 0$, i.e. no outcome is exactly zero –, then we could take logs and use a linear regression. The Tobit model is used if some observations are exactly at the boundary.

First, we derive the likelihood function. Again x_i and β are dropped from the conditioning set for simplicity. The probability of observing $y_i = 0$ is

$$\mathbb{P}[y_i = 0] = \mathbb{P}[y_i^* < 0] = \Phi\left(-\frac{x_i'\beta}{\sigma}\right) .$$

To get the pdf $p(y_i)$ for $y_i > 0$, we first derive the cdf

$$\mathbb{P}[y_i < y] = \mathbb{P}[y_i^* < y] = \mathbb{P}\left[\frac{u_i}{\sigma} < \frac{y_i - x_i'\beta}{\sigma}\right] = \Phi\left(\frac{y_i - x_i'\beta}{\sigma}\right) ,$$

which gives $p(y) = \frac{1}{\sigma}\phi\left(\frac{y - x_i'\beta}{\sigma}\right)$ for $y > 0$. Hence, our observations y_i follow the pdf

$$p(y_i) = \begin{cases} \Phi\left(-\frac{x_i'\beta}{\sigma}\right) & \text{if } y_i = 0 \\ \frac{1}{\sigma}\phi\left(\frac{y - x_i'\beta}{\sigma}\right) & \text{if } y_i > 0 \end{cases} .$$

Let $\mathcal{I} = \{i : y_i = 0\}$ be the set of observations for which $y_i = 0$. Then the log-likelihood function is

$$\ell(\beta, \sigma | Y, X) = \sum_{i=1}^{n} \log p(y_i) = \sum_{i \in \mathcal{I}} \log\left(\Phi\left(-\frac{x_i'}{\sigma}\right)\right) + \sum_{i \notin \mathcal{I}} \log\left(\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right) .$$

To compute partial effects, we need to derive the (conditional) expectation $\mathbb{E}[y_i|x_i]$. Using the result that for $Z \sim N(0,1)$, $\mathbb{E}[Z|Z > c] = \phi(c)/\Phi(c)$ (inverse Miller ratio), we get

$$\mathbb{E}[y_i|y_i > 0] = \mathbb{E}[x_i'\beta + u_i|u_i > -x_i'\beta] = x_i'\beta + \sigma\phi\left(\frac{x_i'\beta}{\sigma}\right)/\Phi\left(\frac{x_i'\beta}{\sigma}\right) .$$

Using this, we obtain

$$\begin{aligned} \mathbb{E}[y_i] &= \mathbb{E}[y_i|y_i = 0]\mathbb{P}[y_i = 0] + \mathbb{E}[y_i|y_i > 0]\mathbb{P}[y_i > 0] \\ &= \mathbb{E}[y_i|y_i > 0]\left(1 - \mathbb{P}[y_i = 0]\right) \\ &= \Phi\left(\frac{x_i'\beta}{\sigma}\right)x_i'\beta + \sigma\phi\left(\frac{x_i'\beta}{\sigma}\right) . \end{aligned}$$

Based on this equation, analogously to the binary outcome case above, we can construct fitted values $\hat{y}_i = \mathbb{E}[y_i|x_i, \hat{\beta}]$, $R^2 = 1 - SSR/SST$ with SSR being the sum of squared residuals $y_i - \hat{y}_i$, and we can compute partial effects by constructing $\mathbb{E}[y_i|x_i = x_2] - \mathbb{E}[y_i|x_i = x_1]$ or $\frac{\partial \mathbb{E}[y_i|x_i = x]}{\partial x}$. Again, only the sign of a coefficient β_j is informative, and again we can test for the partial effect being zero by testing $\mathcal{H}_0 : \beta_j = 0$.

Truncated Model Suppose instead that we do not even observe y_i when $y_i^* < 0$, i.e. we only observe $y_i = y_i^*$ for $y_i^* > 0$. Then we write the likelihood conditional on $y_i^* > 0$. We have

$$p(y_i|y_i^* > 0) = \frac{p(y_i, y_i^* > 0)}{\mathbb{P}[y_i^* > 0]} = \frac{p(y_i^*)\,\mathbf{1}\{y_i^* > 0\}}{\mathbb{P}[y_i^* > 0]} = \frac{1}{\Phi\left[\frac{x_i'\beta}{\sigma}\right]}\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\mathbf{1}\{y_i^* > 0\}\ .$$

This leads to the log-likelihood

$$\ell(\beta, \sigma|Y, X) = \sum_{i \in \mathcal{G}} \log\left(\frac{1}{\sigma}\frac{\phi(\frac{y_i - x_i'\beta}{\sigma})}{\Phi(\frac{x_i'}{\sigma})}\right)\ ,$$

where $\mathcal{G} = \{i : y_i > 0\}$ is the set of observations for which we observe y_i.

4.4.3 Sample Selection Model

Consider a more general setting than in the censored (Tobit) model and let the variables that determine whether $y_i > 0$ or $y_i = 0$ and those that determine the size of y_i if it's positive be different:

$$y_i = d_i y_i^*\ , \quad y_i^* = x_i'\beta + u_i\ , \quad d_i = \mathbf{1}\{d_i^* > 0\}\ , \quad d_i^* = z_i'\gamma + v_i\ , \quad (v_i, \varepsilon_i)' \overset{i.i.d.}{\sim} N(0, I)\ ,$$

where $u_i = \rho v_i + \sigma_\varepsilon \varepsilon_i$. For example, d_i could be the decision on whether a candidate accepted an offered job, y_i^* the wage offered, and y_i the wage observed by researcher. Writing u_i out as $\rho v_i + \sigma_\varepsilon \varepsilon_i$ allows us to distinguish the part of it that is correlated with v_i and the part that is not.

Let $\theta = (\beta', \gamma', \rho, \sigma_\varepsilon)'$. We want to form the (conditional) likelihood

$$\mathcal{L}(\theta|Y, D, Z, X) = \prod_{i=1}^n p(y_i, d_i|x_i, z_i, \theta) = \prod_{i=1}^n p(y_i|d_i, x_i, z_i, \theta)p(d_i|x_i, z_i, \theta)\ .$$

From the Probit model, we know $p(d_i|x_i, z_i, \theta) = \Phi(z_i'\gamma)^{d_i}\Phi(-z_i'\gamma)^{1-d_i}$. We also know that $p(y_i|d_i = 0, x_i, z_i, \theta) = \delta_0(y_i)$ is a point mass at zero. The only missing piece is $p(y_i|d_i = 1, x_i, z_i, \theta) = p(u_i|d_i = 1, x_i, z_i, \theta)|_{u_i = y_i - x_i'\beta}$, for which we need to derive $p(u_i|d_i = 1, x_i, z_i, \theta)$ for $u_i = \rho v_i + \sigma_\varepsilon \varepsilon_i$. For notational simplicity, I drop (x_i, z_i, θ) from the conditioning set. We have

$$p(v_i|d_i = 1) = \frac{p(v_i, d_i = 1)}{p(d_i = 1)} = \frac{p(v_i, v_i \geqslant -z_i'\gamma)}{p(v_i \geqslant -z'\gamma)} = \frac{\phi(v_i)}{\Phi(z_i'\gamma)}\mathbf{1}\{v_i \geqslant -z_i'\gamma\}\ .$$

Also, since v_i and ε_i are independent (given x_i), we can write $p(u_i) = p(\rho v_i + \sigma_\varepsilon \varepsilon_i) =$

$\int p_v(\frac{1}{p}(u_i - \sigma_\varepsilon\varepsilon_i))p_\varepsilon(\varepsilon_i)d\varepsilon_i$. Conditional on $d_i = 1$, this yields

$$p_u(u_i|d_i = 1) = \frac{1}{\Phi(z_i\gamma')} \int \mathbf{1}\left\{\varepsilon_i \leqslant \frac{1}{\sigma_\varepsilon}(u_i + \rho z_i'\gamma)\right\} \phi\left(\frac{1}{p}(u_i - \sigma_\varepsilon\varepsilon_i)\right)\phi(\varepsilon_i)d\varepsilon_i \ .$$

After a little algebra, we see that

$$\begin{aligned}\phi\left(\frac{1}{p}(u_i - \sigma_\varepsilon\varepsilon_i)\right)\phi(\varepsilon_i) =& \rho(\sigma_\varepsilon^2 + \rho^2)^{-1/2}\phi\left(\frac{1}{(\sigma_\varepsilon^2+\rho^2)^{1/2}}u_i\right)\left(\frac{\rho^2}{\sigma_\varepsilon^2+\rho^2}\right)^{1/2} \\ & \phi\left(\frac{1}{(\sigma_\varepsilon^2+\rho^2)^{1/2}}\left(\varepsilon_i - \frac{\sigma_\varepsilon u_i}{\sigma_\varepsilon^2+\rho^2}\right)\right) \ ,\end{aligned}$$

and, in turn,

$$\begin{aligned}p(u_i|d_i = 1) =& \frac{1}{\Phi(-z_i'\gamma)}\rho(\sigma_\varepsilon^2+\rho^2)^{-\frac{1}{2}}\phi\left(\frac{1}{(\sigma_\varepsilon^2+\rho^2)^{\frac{1}{2}}}u_i\right) \\ & \int \mathbf{1}\left\{\varepsilon_i \leqslant \frac{1}{\sigma_\varepsilon}(u_i+\rho z_i'\gamma)\right\}\left(\frac{\rho^2}{\sigma_\varepsilon^2+\rho^2}\right)^{-\frac{1}{2}}\phi\left((\frac{\rho^2}{\sigma_\varepsilon^2+\rho^2})^{-\frac{1}{2}}(\varepsilon_i - \frac{\sigma_\varepsilon u_i}{\sigma_\varepsilon^2+\rho^2})\right)d\varepsilon_i \\ =& \frac{1}{\Phi(-z_i'\gamma)}\rho(\sigma_\varepsilon^2+\rho^2)^{-\frac{1}{2}}\phi\left(\frac{1}{(\sigma_\varepsilon^2+\rho^2)^{\frac{1}{2}}}u_i\right)\Phi\left(\left(\frac{\rho^2}{\sigma_\varepsilon^2+\rho^2}\right)^{-\frac{1}{2}}\left(\frac{1}{\sigma_\varepsilon}(u_i+\rho z_i'\gamma) - \frac{\sigma_\varepsilon u_i}{\sigma_\varepsilon^2+\rho^2}\right)\right)\end{aligned}$$

Putting all pieces together, we get

$$\mathcal{L}(\theta|Y,D,Z,X) = \prod_{i=1}^{n}\Phi(z_i'\gamma)^{d_i}\Phi(-z_i'\gamma)^{1-d_i}\left[p(u_i|d_i=1,x_i,z_i,\theta)|_{u_i=y_i-x_i'\beta}\right]^{d_i} \ ,$$

with $p(u_i|d_i = 1, x_i, z_i, \theta)$ given by the expression above.

Heckman two-step procedure ("Heckit") One can also get an estimator for θ in the sample selection model using a two-step procedure. Note that

$$\begin{aligned}\mathbb{E}[y_i|d_i = 1] &= \mathbb{E}[y_i^*|d_i = 1] \\ &= x_i'\beta + \mathbb{E}[u_i|v_i \geqslant -z_i'\gamma] \\ &= x_i'\beta + \rho\mathbb{E}[v_i|v_i \geqslant -z_i'\gamma] \\ &= x_i'\beta + \rho\lambda(z_i'\gamma) \ ,\end{aligned}$$

where $\lambda(x) = \frac{\phi(x)}{\Phi(x)}$ is the inverse Miller ratio. The idea of this method is to use the observations with $d_i = 1$ to estimate (β, ρ) by regressing y_i on x and $\lambda(z_i'\gamma)$. First, a probit model is estimated for d_i with covariates z_i. This gives an estimate for γ and yields $\hat{\lambda}_i = \lambda(z_i'\hat{\gamma})$.

Then, one uses the observations with $d_i = 1$ to regress y_i on x_i and $\hat{\lambda}_i$ and obtain $\hat{\beta}$ and $\hat{p}$.

This method is simpler than (deriving and) maximizing the likelihood above. However, it is also less efficient by virtue of being a two-step procedure and because it uses only observations with $d_i = 1$ in the second step. Also, the standard errors of $(\hat{\beta}, \hat{p})$ need to be adjusted for the fact that λ is estimated with $\hat{\lambda}$.

4.5 Bayesian Analysis of the Linear Regression Model

This section discusses Bayesian estimation of the linear regression model. Under appropriate (and commonly used) distributional families for the likelihood and prior, the posterior can be derived analytically. This illustrates some fundamental aspects of Bayesian estimation that apply for more general models (likelihood and prior specifications) – discussed in turn in Section 4.6 – as well.

4.5.1 Estimating $\beta|\sigma^2$

As laid out in Section 4.1 above, the linear regression model under conditional Normality of error terms leads to the likelihood function

$$p(Y|\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} exp\left\{ -\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) \right\} .$$

For now, assume we are interested only in estimating β and take σ^2 as given. Suppose our prior is Normal as well: $\beta|\sigma^2 \sim N(\underline{\beta}, \sigma^2\underline{V})$.[11] For simplicity, the conditioning on σ^2 is dropped from the following expressions. The posterior of β is then[12]

$$\begin{aligned} p(\beta|Y) &\propto p(Y|\beta)p(\beta) \\ &\propto exp\left\{ -\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) \right\} exp\left\{ -\frac{1}{2\sigma^2}(\beta - \underline{\beta})'\underline{V}^{-1}(\beta - \underline{\beta}) \right\} \\ &\propto exp\left\{ -\frac{1}{2\sigma^2}\left[-\beta'X'Y - Y'X\beta + \beta'X'X\beta + \beta'\underline{V}^{-1}\beta - \beta'\underline{V}^{-1}\underline{\beta} - \underline{\beta}'\underline{V}^{-1}\beta \right] \right\} \\ &\propto exp\left\{ -\frac{1}{2\sigma^2}\left[\beta'[X'X + \underline{V}^{-1}]\beta + -2(Y'X + \underline{\beta}'\underline{V}^{-1})\beta \right] \right\} . \end{aligned}$$

[11] Conditioning the prior on σ^2 is natural because the meaning of size of coefficients is only obtained given σ^2 (it determines the scale of data).

[12] More precisely, this is the conditional posterior of $\beta|\sigma^2$.

λ. Model selection and model averaging are two ways to deal with this.[31] Under the former we select the λ that maximizes the MDD $p(Y|\lambda)$ (recall that in Section 4.5.1.1, we argued that the λ^* that maximizes the MDD $p(Y|\lambda)$ leads to the model that optimizes the trade-off between in-sample fit and model complexity, leading to the best out-of-sample fit), while under the latter we consider many values for λ and construct a model-averaged posterior as the weighted average of the posteriors under different values for λ.

Both model selection and model averaging are facilitated by hierarchical Bayes modeling. Usually, a Uniform hyperprior is assumed: $p(\lambda) \propto c$. This means that we regard each value of λ (each model) as equally likely a-priori. We get the prior $p(\theta, \lambda) = p(\theta|\lambda)p(\lambda) \propto p(\theta|\lambda)$. Under this prior, the marginal posterior of λ has the same shape as the log MDD conditional on λ, $p(Y|\lambda)$:

$$p(\lambda|Y) \propto p(Y|\lambda)p(\lambda) \propto p(Y|\lambda) \; .$$

Therefore, the λ^* that maximizes $p(Y|\lambda)$ is equal to the mode of $p(\lambda|Y)$ in the hierarchical Bayes model, just as in the simple example above.[32] Moreover, averaging $p(\theta|Y, \lambda)$ for many values λ using weights proportional to $p(Y|\lambda_j)p(\lambda_j)$ is equivalent to taking the marginal posterior

$$p(\theta|Y) = \int p(\theta, \lambda|Y)p(\lambda|Y)d\lambda \; ,$$

analogous to the simple example above.[33] [34]

Sometimes, we can find λ^* and $p(\theta|Y)$ analytically. In other cases, we find them numerically. Given a numerical approximation of the posterior $p(\theta, \lambda|Y)$, $\{(\theta^m, \lambda^m), W^m\}_{m=1}^M$, we get a numerical approximation of the marginal posterior $p(\lambda|Y)$ by simply taking $\{\lambda^m, W^m\}_{m=1}^M$. The mode of this distribution is λ^*. Similarly, we get a numerical approximation of the marginal posterior $p(\theta|Y)$ by simply taking $\{\theta^m, W^m\}_{m=1}^M$.

Appendix

Definition 34. *A statistic $T(Y)$ is a sufficient statistic for a parameter β if conditional an $T(Y)$ the distribution of the data Y does not depend an β.*

[31] The third is to conduct a prior sensitivity analysis, i.e. to explore the robustness of results to changes in the prior (in the values of (some) hyperparameters).

[32] Note that this procedure works always, whereas maximizing $p(Y|\lambda)$ w.r.t. λ is cumbersome and inefficient if $p(Y|\lambda)$ is not available analytically. One would first need to derive $p(\theta|Y, \lambda)$ and $p(Y|\lambda)$ many times, each time for a different value of λ. Then, among all those values, one would need to choose the value of λ that leads to the highest $p(Y|\lambda)$.

[33] In the example, we average $p(\theta|Y, \lambda_j)$ for two discrete values $\lambda_j \in \{0, 1\}$ using weights $\pi_{j,n}$ proportional to $p(Y|\lambda_j)p(\lambda_j)$, whereas here we perform an integration as we average over a continuous λ.

[34] See Giannone et al. (2015) for a more detailed discussion on hierarchical Bayes modeling.

Proposition 24 (Factorization Theorem).
Given $p(Y|\beta)$, $T(Y)$ is a sufficient statistic for β iff there are functions $g(T(Y)|\beta)$ and $h(y)$ s.t. we can write $p(Y|\beta) = g(T(Y)|\beta)h(Y)$.

The following shows that in the linear regression model, under Normality of $u_i|x_i$, $T(Y) = \hat{\beta} = (X'X)^{-1}X'Y$ is a sufficient statistic for β. This means that an experimenter who has observed $T(Y) = \hat{\beta}$ has as much information about β as one who has observed the whole sample Y and X.

$$\begin{aligned}(Y - X\beta)'(Y - X\beta) &= (P_XY + M_XY - X\beta)'(P_XY + M_XY - X\beta) \\ &= (M_XY)'M_XY + (P_XY - X\beta)'(P_XY - X\beta) + \underbrace{2M_XY(P_XY - X\beta)}_{=0 \text{ as } M_XP_X=0 \text{ and } M_XX=0} .\end{aligned}$$

Note that $P_XY = X\hat{\beta}$. As a result, we can write

$$p(y|\beta) = (2\pi\sigma^2)^{-(n-k)/2} \exp\left\{-\frac{1}{2\sigma^2}(M_Xy)'M_XY\right\} (2\pi\sigma^2)^{-k/2} \exp\left\{-\frac{1}{2\sigma^2}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)\right\} .$$

This means that $T(Y) = (X'X)^{-1}X'Y = \hat{\beta}$ is a sufficient statistic for β.

Moreover, from this expression, we can see that $\hat{\beta}|\beta \sim N(\beta, \sigma^2(X'X)^{-1})$. Thus, we can treat our inference problem on β as an experiment of drawing a $(k \times 1)$ RV $\hat{\beta}$ with distribution $\hat{\beta}|\beta \sim N(\beta, \sigma^2(X'X)^{-1})$ and then estimating β, rather than drawing a whole sample of Y (and X) and estimating the parameter β that appears in the linear regression model. Note that the former is exactly what we have done throughout Chapter 2.

Derivation of ML Approach from the Kullback-Leibler Inequality Let $\mathcal{Y}$ denote the support of y_i given x_i: $\mathcal{Y} = \{y : p_o(y|x) > 0\}$. Here, $p_0(y|x) = f(y|x, \theta_0)$ is the correct density of $y|x$. By Jensen's inequality for concave functions, we have $\mathbb{E}[log(y)|x] \leq log(\mathbb{E}[y|x])$. Analogously, for any (other) distribution $g(y|x, \theta)$, we have

$$\int_{\mathcal{Y}} log\left(\frac{g(y|x)}{p_o(y|x)}\right) p_o(y|x)dy \leq log\left(\int_{\mathcal{Y}} \frac{g(y|x)}{p_o(y|x)} p_o(y|x)dy\right) = log\left(\int_{\mathcal{Y}} g(y|x)dy\right) = log(1) = 0 .$$

The dependence on θ is suppressed for notational simplicity. Realising that $log(\frac{g(y|x)}{p_o(y|x)}) = -log(\frac{p_o(y|x)}{g(y|x)})$ leads to the Kullback-Leibler Inequality:

$$\int_y log\left(\frac{p_o(y|x)}{g(y|x)}\right) p_o(y|x) dy \geq 0 \ .$$

Rearranging gives $E[log(p_o(y|x))|x] \geq E[log(g(y|x))|x]$. This holds with an equality if $g(y|x) = p_0(y|x)$, i.e. if $g(y|x, \theta) = f(y|x, \theta_0)$; we picked the right distributional assumption and the right value for parameters θ_0 that index this distribution. Assuming that we picked the right distribution, we have

$$E[log(f(y|x, \theta_0))|x] > \mathbb{E}[log(f(y|x, \theta))|x] \quad \forall\, \theta \in \Theta \ , \ \theta \neq \theta_o \ \ ,$$

and therefore, $\theta_o = \arg\max_{\theta\in\Theta} \mathbb{E}[l_i(\theta)|x]$ where $l_i(\theta) = log(f(y|x_i, \theta))$ is the conditional log likelihood of observation i.

Proof of Information Matrix Equality Under Scalar Parameters Suppose θ is a scalar and $p(y|\theta)$ is sufficiently smooth s.t. we can exchange the order of integration and differentiation. The score is $s(\theta) = \frac{\partial \log p(Y|\theta)}{\partial \theta}$ and the Hessian is $H(\theta) = \frac{\partial s(\theta)}{\partial \theta} = \frac{p''(Y|\theta)p(Y|\theta) - p'(Y|\theta)^2}{p(Y|\theta)^2}$. However, the proof holds likewise for $s_i(\theta) = \frac{\partial \log p(y_i|\theta)}{\partial \theta}$ and $H_i(\theta) = \frac{\partial s_i(\theta)}{\partial \theta}$.

First, note that

$$\mathbb{E}[s(\theta_0)|\theta_0] = \int \frac{p'(Y|\theta_0)}{p(Y|\theta_0)} p(Y|\theta_0) dY = \int p'(Y|\theta_0) dY = \frac{\partial}{\partial \theta}\left[\int p(Y|\theta) dY\right]_{\theta=\theta_0} = 0 \ ,$$

because $\int p(Y|\theta)dY = 1$ and its derivative w.r.t. θ is zero. Similarly,

$$\mathbb{E}[s(\theta_0)^2|\theta_0] = \int \left(\frac{p'(Y|\theta_0)}{p(Y|\theta_0)}\right)^2 p(Y|\theta_0) dY = \int \frac{p'(Y|\theta_0)^2}{p(Y|\theta_0)} dY \ .$$

Using this result, we get the information matrix equality:

$$-\mathbb{E}[H(\theta_0)|\theta_0] = -\int H(\theta_0) p(Y|\theta_0) dY = -\int \frac{p''(Y|\theta_0)p(Y|\theta_0) - p'(Y|\theta_0)^2}{p(Y|\theta_0)} dY = \mathbb{E}[s(\theta_0)^2|\theta_0] \ .$$

This holds because $\int p''(Y|\theta_0)dY = \int \frac{\partial^2}{\partial\theta\partial\theta'} p(Y|\theta)|_{\theta=\theta_0} dY = 0$, as we can once again exchange the order of integration and differentiation, and we have $\int p(Y|\theta_0)dY = 1$.

Proof of Cramer-Rao Lower Bound Under Scalar Parameters As before, suppose θ is a scalar and $p(y|\theta)$ is sufficiently smooth s.t. we can exchange the order of integration and differentiation. Let $\tilde{\theta}(Y)$ be an unbiased estimator for θ_0, i.e. $\mathbb{E}[\tilde{\theta}(Y)|\theta_0] = \int \tilde{\theta}(Y)p(Y|\theta_0)dY = \theta_0$. Consider the matrix

$$\Omega \equiv \mathbb{V}\begin{bmatrix} \tilde{\theta}(Y) - \theta_0 \\ s(\theta_0) \end{bmatrix} = \begin{bmatrix} \mathbb{V}[\tilde{\theta}] & \mathbb{E}[\tilde{\theta}(Y)s(\theta_0)] \\ \mathbb{E}[\tilde{\theta}(Y)s(\theta_0)] & I(\theta_0) \end{bmatrix} ,$$

whereby we used the fact that $\mathbb{E}[(\tilde{\theta}(Y) - \theta_0)s(\theta_0)] = \mathbb{E}[\tilde{\theta}(Y)s(\theta_0)]$ because $\mathbb{E}[s(\theta_0)] = 0$. We have

$$\mathbb{E}[\tilde{\theta}(Y)s(\theta_0)] = \int \tilde{\theta}(Y)\frac{p'(yY|\theta_0)}{p(Y|\theta_0)}p(Y|(\theta_0))dY = \frac{\partial}{\partial\theta}\int \tilde{\theta}(Y)p(Y|\theta)dY|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}\theta|_{\theta=\theta_0} = 1 .$$

Take $X = \begin{bmatrix} 1 \\ I^{-1}(\theta_0) \end{bmatrix}$. Because Ω is p.s.d., we know

$$X'\Omega X = \begin{bmatrix} 1 - I^{-1}(\theta_0) \end{bmatrix} \begin{bmatrix} \mathbb{V}[\tilde{\theta}] & 1 \\ 1 & I(\theta_0) \end{bmatrix} \begin{bmatrix} 1 \\ -I^{1}(\theta_0) \end{bmatrix} = \mathbb{V}[\tilde{\theta}] - I^{-1}(\theta_0) \geqslant 0 ,$$

which implies $V[\tilde{\theta}] \geqslant I^{-1}(\theta_0)$.