4
Loss functions and estimation

In the previous chapter we reviewed some results of linear (least squares) models without making the loss function explicit. In this chapter we remedy this and extend the discussion to various other (sometimes referred to as "robust") approaches. That the loss function determines the properties of estimators is common to classical and Bayesian statistics (whether made explicit or not). We'll review a few loss functions and the associated expected loss minimizing estimators. Then we briefly review maximum likelihood estimation (*MLE*) and nonlinear regression.

4.1 Loss functions

Let the loss function associated with the estimator $\hat{\theta}$ for θ be $C\left(\hat{\theta}, \theta\right)$ and the posterior distribution function be $f\left(\theta \mid y\right)$,[1] then minimum expected loss is

$$\min_{\hat{\theta}} E\left[C\left(\hat{\theta}, \theta\right)\right] = \int C\left(\hat{\theta}, \theta\right) f\left(\theta \mid y\right) d\theta$$

Briefly, a symmetric quadratic loss function results in an estimator equal to the posterior mean, a linear loss function results in an estimator equal to a quantile of the posterior distribution $f\left(\theta \mid y\right)$, and an all or nothing loss function results in an estimator for θ equal to the posterior mode.

[1] A source of controversy is whether the focus is the posterior distribution $f\left(\theta \mid y\right)$ or the likelihood function $f\left(y \mid \theta\right)$; see Poirier [1995]. We initially focus on the posterior distribution then review *MLE*.

4.1.1 Quadratic loss

The quadratic loss function is

$$C\left(\hat{\theta},\theta\right)=\begin{array}{ll} c_1\left(\hat{\theta}-\theta\right)^2 & \hat{\theta}\leq\theta \\ c_2\left(\hat{\theta}-\theta\right)^2 & \hat{\theta}>\theta \end{array}$$

First order conditions are

$$\frac{d}{d\hat{\theta}}\left\{\begin{array}{c} \int_{\hat{\theta}}^{\infty} c_1\left(\hat{\theta}-\theta\right)^2 f\left(\theta\mid y\right)d\theta \\ +\int_{-\infty}^{\hat{\theta}} c_2\left(\hat{\theta}-\theta\right)^2 f\left(\theta\mid y\right)d\theta \end{array}\right\}=0$$

Rearrangement produces

$$\frac{d}{d\hat{\theta}}\left\{\begin{array}{c} c_1\left(1-F\left(\hat{\theta}\right)\right)\hat{\theta}^2-2c_1\hat{\theta}\int_{\hat{\theta}}^{\infty}\theta f\left(\theta\mid y\right)d\theta \\ +c_1\int_{\hat{\theta}}^{\infty}\theta^2 f\left(\theta\mid y\right)d\theta+c_2F\left(\hat{\theta}\right)\hat{\theta}^2 \\ -2c_2\int_{-\infty}^{\hat{\theta}}\theta f\left(\theta\mid y\right)d\theta+c_2\int_{-\infty}^{\hat{\theta}}\theta^2 f\left(\theta\mid y\right)d\theta \end{array}\right\}=0$$

where $F\left(\hat{\theta}\right)$ is the cumulative posterior distribution function for θ given the data y evaluated at $\hat{\theta}$. Differentiation reveals

$$\left\{\begin{array}{c} c_1\left[\begin{array}{c} 2\hat{\theta}\left(1-F\left(\hat{\theta}\right)\right)-\hat{\theta}^2 f\left(\hat{\theta}\right) \\ -2\int_{\hat{\theta}}^{\infty}\theta f\left(\theta\mid y\right)d\theta+2\hat{\theta}^2 f\left(\hat{\theta}\right)-\hat{\theta}^2 f\left(\hat{\theta}\right) \end{array}\right] \\ +c_2\left[\begin{array}{c} 2\hat{\theta}F\left(\hat{\theta}\right)+\hat{\theta}^2 f\left(\hat{\theta}\right) \\ -2\int_{-\infty}^{\hat{\theta}}\theta f\left(\theta\mid y\right)d\theta-2\hat{\theta}^2 f\left(\hat{\theta}\right)+\hat{\theta}^2 f\left(\hat{\theta}\right) \end{array}\right] \end{array}\right\}=0$$

Simplification yields

$$\begin{array}{rl} & \hat{\theta}\left[c_1\left(1-F\left(\hat{\theta}\right)\right)+c_2F\left(\hat{\theta}\right)\right] \\ = & c_1\left(1-F\left(\hat{\theta}\right)\right)E\left[\theta\mid y,\hat{\theta}\leq\theta\right]+c_2F\left(\hat{\theta}\right)E\left[\theta\mid y,\hat{\theta}>\theta\right] \end{array}$$

Or,

$$\hat{\theta}=\frac{c_1\left(1-F\left(\hat{\theta}\right)\right)E\left[\theta\mid y,\theta\geq\hat{\theta}\right]+c_2F\left(\hat{\theta}\right)E\left[\theta\mid y,\theta<\hat{\theta}\right]}{c_1\left(1-F\left(\hat{\theta}\right)\right)+c_2F\left(\hat{\theta}\right)}$$

In other words, the quadratic expected loss minimizing estimator for θ is a cost-weighted average of truncated means of the posterior distribution. If $c_1=c_2$ (symmetric loss), then $\hat{\theta}=E\left[\theta\mid y\right]$, the mean of the posterior distribution.

4.1.2 Linear loss

The linear loss function is

$$C\left(\hat{\theta},\theta\right)=\begin{array}{ll} c_1\left|\hat{\theta}-\theta\right| & \hat{\theta}\leq\theta \\ c_2\left|\hat{\theta}-\theta\right| & \hat{\theta}>\theta \end{array}$$

First order conditions are

$$\frac{d}{d\hat{\theta}}\left\{\begin{array}{c} -\int_{\hat{\theta}}^{\infty} c_1\left(\hat{\theta}-\theta\right) f\left(\theta\mid y\right) d\theta \\ +\int_{-\infty}^{\hat{\theta}} c_2\left(\hat{\theta}-\theta\right) f\left(\theta\mid y\right) d\theta \end{array}\right\}=0$$

Rearranging yields

$$\frac{d}{d\hat{\theta}}\left\{\begin{array}{c} -c_1\hat{\theta}\left(1-F\left(\hat{\theta}\right)\right)+c_1\int_{\hat{\theta}}^{\infty}\theta f\left(\theta\mid y\right) d\theta \\ +c_2\hat{\theta}F\left(\hat{\theta}\right)-c_2\int_{-\infty}^{\hat{\theta}}\theta f\left(\theta\mid y\right) d\theta \end{array}\right\}=0$$

Differentiation produces

$$\begin{aligned} 0 &= c_1\left[-\left(1-F\left(\hat{\theta}\right)\right)+\hat{\theta}f\left(\hat{\theta}\right)-\hat{\theta}f\left(\hat{\theta}\right)\right] \\ &\quad +c_2\left[F\left(\hat{\theta}\right)+\hat{\theta}f\left(\hat{\theta}\right)-\hat{\theta}f\left(\hat{\theta}\right)\right] \end{aligned}$$

Simplification reveals

$$c_1\left(1-F\left(\hat{\theta}\right)\right)=c_2F\left(\hat{\theta}\right)$$

Or

$$F\left(\hat{\theta}\right)=\frac{c_1}{c_1+c_2}$$

The expected loss minimizing estimator is the quantile that corresponds to the relative cost $\frac{c_1}{c_1+c_2}$. If $c_1=c_2$, then the estimator is the median of the posterior distribution.

4.1.3 All or nothing loss

The all or nothing loss function is

$$C\left(\hat{\theta},\theta\right)=\begin{array}{ll} c_1 & \hat{\theta}<\theta \\ 0 & \hat{\theta}=\theta \\ c_2 & \hat{\theta}>\theta \end{array}$$

If $c_1>c_2$, then we want to choose $\hat{\theta}>\theta$, so $\hat{\theta}$ is the upper limit of support for $f\left(\theta\mid y\right)$. If $c_1<c_2$, then we want to choose $\hat{\theta}<\theta$, so $\hat{\theta}$ is the lower limit of

support for $f(\theta \mid y)$. If $c_1 = c_2$, then we want to choose $\hat{\theta}$ to maximize $f(\theta \mid y)$, so $\hat{\theta}$ is the mode of the posterior distribution.[2]

4.2 Nonlinear Regression

Many accounting and business settings call for analysis of data involving limited dependent variables (such as discrete choice models discussed in the next chapter).[3] Nonlinear regression frequently complements our understanding of standard maximum likelihood procedures employed for estimating such models as well as providing a means for addressing alternative functional forms. Here we review some basics of nonlinear least squares including Newton's method of optimization, Gauss-Newton regression (*GNR*), and artificial regressions.

Our discussion revolves around minimizing a smooth, twice continuously differentiable function, $Q(\beta)$. It's convenient to think $Q(\beta)$ equals $SSR(\beta)$, the residual sum of squares, but $-Q(\beta)$ might also refer to maximization of the log-likelihood.

4.2.1 Newton's method

A second order Taylor series approximation of $Q(\beta)$ around some initial values for β, say $\beta_{(0)}$ yields

$$Q^*(\beta) = Q\left(\beta_{(0)}\right) + g_{(0)}^T\left(\beta - \beta_{(0)}\right) + \frac{1}{2}\left(\beta - \beta_{(0)}\right)^T H_{(0)}\left(\beta - \beta_{(0)}\right)$$

where $g(\beta)$ is the $k \times 1$ gradient of $Q(\beta)$ with typical element $\frac{\partial Q(\beta)}{\partial \beta_j}$, $H(\beta)$ is the $k \times k$ Hessian of $Q(\beta)$ with typical element $\frac{\partial^2 Q(\beta)}{\partial \beta_j \partial \beta_i}$, and for notational simplicity, $g_{(0)} \equiv g\left(\beta_{(0)}\right)$ and $H_{(0)} \equiv H\left(\beta_{(0)}\right)$. The first order conditions for a minimum of $Q^*(\beta)$ with respect to β are

$$g_{(0)} + H_{(0)}\left(\beta - \beta_{(0)}\right) = 0$$

Solving for β yields a new value

$$\beta_{(1)} = \beta_{(0)} - H_{(0)}^{-1} g_{(0)}$$

This is the core of Newton's method. Successive values $\beta_{(1)}, \beta_{(2)}, \ldots$ lead to an approximation of the global minimum of $Q(\beta)$ at $\hat{\beta}$. If $Q(\beta)$ is approximately

[2]For a discrete probability mass distribution, the optimal estimator may be either the limit of support or the mode depending on the difference in cost. Clearly, large cost differentials are aligned with the limits and small cost differences are aligned with the mode.

[3]This section draws heavily from Davidson and MacKinnon [1993].

quadratic, as applies to sums of squares when sufficiently close to their minima, Newton's method usually converges quickly.[4]

4.2.2 Gauss-Newton regression

When minimizing a sum of squares function it is convenient to write the criterion as

$$Q(\beta) = \frac{1}{n} SSR(\beta) = \frac{1}{n} \sum_{t=1}^{n} (y_t - x_t(\beta))^2$$

Now, explicit expressions for the gradient and Hessian can be found. The gradient for the i^{th} element is

$$g_i(\beta) = -\frac{2}{n} \sum_{t=1}^{n} X_{ti}(\beta)(y_t - x_t(\beta))$$

where $X_{ti}(\beta)$ is the partial derivative of $x_t(\beta)$ with respect to β_i. The more compact matrix notation is

$$\mathbf{g}(\beta) = -\frac{2}{n} \mathbf{X}^T(\beta)(\mathbf{y} - \mathbf{x}(\beta))$$

The Hessian $H(\beta)$ has typical element

$$H_{ij}(\beta) = -\frac{2}{n} \sum_{t=1}^{n} (y_t - x_t(\beta)) \frac{\partial X_{ti}(\beta)}{\partial \beta_j} - X_{ti}(\beta) X_{tj}(\beta)$$

Evaluated at β_0, this expression is asymptotically equivalent to[5]

$$\frac{2}{n} \sum_{t=1}^{n} X_{ti}(\beta) X_{tj}(\beta)$$

In matrix notation this is

$$D(\beta) = \frac{2}{n} X^T(\beta) X(\beta)$$

and $D(\beta)$ is positive definite when $X(\beta)$ is full rank. Now, writing Newton's method as

$$\beta_{(j+1)} = \beta_{(j)} - D_{(j)}^{-1} g_{(j)}$$

---

[4] If $Q^*(\beta)$ is strictly convex, as it is if and only if the Hessian is positive definite, then $\beta_{(1)}$ is the global minimum of $Q^*(\beta)$. Please consult other sources, such as Davidson and MacKinnon [2003, ch. 6] and references therein, for additional discussion of Newton's method including search direction, step size, and stopping rules.

[5] Since $y_t = x_t(\beta_0) + u_t$, the first term becomes $-\frac{2}{n} \sum_{t=1}^{n} \frac{\partial X_{ti}(\beta)}{\partial \beta_j} u_t$. By the law of large numbers this term tends to 0 as $n \to \infty$.

and substituting the above results we have the classic Gauss-Newton result

$$\begin{aligned}\beta_{(j+1)} &= \beta_{(j)} - \left(\frac{2}{n}X_{(j)}^T X_{(j)}\right)^{-1}\left(-\frac{2}{n}\mathbf{X}_{(j)}^T\left(\mathbf{y}-\mathbf{x}_{(j)}\right)\right) \\ &= \beta_{(j)} + \left(\mathbf{X}_{(j)}^T\mathbf{X}_{(j)}\right)^{-1}\mathbf{X}_{(j)}^T\left(\mathbf{y}-\mathbf{x}_{(j)}\right)\end{aligned}$$

Artificial regression

The second term can be readily estimated by an artificial regression. It's called an artificial regression because functions of the variables and model parameters are employed. This artificial regression is referred to as a Gauss-Newton regression (*GNR*)

$$\mathbf{y}-\mathbf{x}\left(\beta\right) = \mathbf{X}\left(\beta\right)b + \ residuals$$

To be clear, Gaussian projection (*OLS*) produces the following estimate

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T\left(\beta\right)\mathbf{X}\left(\beta\right)\right)^{-1}\mathbf{X}^T\left(\beta\right)\left(\mathbf{y}-\mathbf{x}\left(\beta\right)\right)$$

To appreciate the *GNR*, consider a linear regression where X is the matrix of regressors. Then $\mathbf{X}\left(\beta\right)$ is simply replaced by X, the *GNR* is

$$\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{(0)} = \mathbf{X}\mathbf{b} + \ residuals$$

and the artificial parameter estimates are

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{(0)}\right) = \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{(0)}$$

where $\hat{\boldsymbol{\beta}}$ is the *OLS* estimate. Rearranging we see that the Gauss-Newton estimate replicates *OLS*, $\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(0)} + \hat{\mathbf{b}} = \boldsymbol{\beta}_{(0)} + \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{(0)} = \hat{\boldsymbol{\beta}}$, as expected.

Covariance matrices

Return to the *GNR* above and substitute the nonlinear parameter estimates

$$\mathbf{y}-\mathbf{x}\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{X}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{b} + residuals$$

The artificial regression estimate is

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T\left(\hat{\boldsymbol{\beta}}\right)\mathbf{X}\left(\hat{\boldsymbol{\beta}}\right)\right)^{-1}\mathbf{X}^T\left(\hat{\boldsymbol{\beta}}\right)\left(\mathbf{y}-\mathbf{x}\left(\hat{\boldsymbol{\beta}}\right)\right)$$

Since the first order or moment conditions require

$$\mathbf{X}^T\left(\hat{\boldsymbol{\beta}}\right)\left(\mathbf{y}-\mathbf{x}\left(\hat{\boldsymbol{\beta}}\right)\right) = 0$$

this regression cannot have any explanatory power, $\hat{\mathbf{b}} = \mathbf{0}$. Though this may not seem very interesting, it serves two useful functions. First, it provides a check on

the consistency of the nonlinear optimization routine. Second, as it is the *GNR* variance estimate, it provides a quick estimator of the covariance matrix for the parameter estimates

$$\widehat{Var}\left[\hat{b}\right] = s^2\left(\mathbf{X}^T\left(\hat{\boldsymbol{\beta}}\right)\mathbf{X}\left(\hat{\boldsymbol{\beta}}\right)\right)^{-1}$$

and it is readily available from the artificial regression.

Further, this same *GNR* readily supplies a heteroskedastic-consistent covariance matrix estimator. If $E\left[uu^T\right] = \Omega$, then a heteroskedastic-consistent covariance matrix estimator is

$$\widehat{Var}\left[\hat{\mathbf{b}}\right] = \left(\mathbf{X}^T\left(\hat{\boldsymbol{\beta}}\right)\mathbf{X}\left(\hat{\boldsymbol{\beta}}\right)\right)^{-1}\mathbf{X}^T\left(\hat{\boldsymbol{\beta}}\right)\hat{\Omega}\mathbf{X}\left(\hat{\boldsymbol{\beta}}\right)\left(\mathbf{X}^T\left(\hat{\boldsymbol{\beta}}\right)\mathbf{X}\left(\hat{\boldsymbol{\beta}}\right)\right)^{-1}$$

where $\hat{\Omega}$ is a diagonal matrix with t^{th} element equal to the squared residual u_t^2. Next, we turn to maximum likelihood estimation and exploit some insights gained from nonlinear regression as they relate to typical *MLE* settings.

4.3 Maximum likelihood estimation (MLE)

Maximum likelihood estimation (*MLE*) applies to a wide variety of problems.[6] Since it is the most common method for estimating discrete choice models and discrete choice models are central to the discussion of accounting choice, we focus the discussion of *MLE* around discrete choice models.

4.3.1 Parameter estimation

The most common method for estimating the parameters of discrete choice models is maximum likelihood. Recall the likelihood is defined as the joint density for the parameters of interest β conditional on the data X_t. For binary choice models and $Y_t = 1$ the contribution to the likelihood is $F\left(X_t\beta\right)$, and for $Y_t = 0$ the contribution to the likelihood is $1 - F\left(X_t\beta\right)$ where these are combined as binomial draws. Hence,

$$L\left(\beta|X\right) = \prod_{t=1}^{n} F\left(X_t\beta\right)^{Y_t}\left[1 - F\left(X_t\beta\right)\right]^{1-Y_t}$$

The log-likelihood is

$$\ell\left(\beta|X\right) \equiv logL\left(\beta|X\right) = \sum_{t=1}^{n} Y_t log\left(F\left(X_t\beta\right)\right) + \left(1 - Y_t\right) log\left(1 - F\left(X_t\beta\right)\right)$$

[6] This section draws heavily from Davidson and MacKinnon [1993], chapter 8.

Since this function for binary response models like probit and logit is globally concave, numerical maximization is straightforward. The first order conditions for a maximum are

$$\sum_{t=1}^{n} \frac{Y_t f(X_t\beta)X_{it}}{F(X_t\beta)} - \frac{(1-Y_t)f(X_t\beta)X_{ti}}{1-F(X_t\beta)} = 0 \quad i = 1, \ldots, k$$

where $f(\cdot)$ is the density function. Simplifying yields

$$\sum_{t=1}^{n} \frac{[Y_t - F(X_t\beta)]f(X_t\beta)X_{ti}}{F(X_t\beta)[1-F(X_t\beta)]} = 0 \quad i = 1, \ldots, k$$

For the logit model the first order conditions simplify to

$$\sum_{t=1}^{n} \left[Y_t - \Lambda\left(X_{ti}\right)\right] X_{ti} = 0 \quad i = 1, \ldots, k$$

since the logit density is $\lambda\left(X_{ti}\right) = \Lambda\left(X_{ti}\right)\left[1 - \Lambda\left(X_{ti}\right)\right]$ where $\Lambda(\cdot)$ is the logit (cumulative) distribution function.

Notice the above first order conditions look like the first order conditions for weighted nonlinear least squares with weights given by $\left[F\left(1-F\right)\right]^{-1/2}$. This is sensible because the error term in the nonlinear regression

$$Y_t = F\left(X_t\beta\right) + \varepsilon_t$$

has mean zero and variance

$$\begin{aligned} E\left[\varepsilon_t^2\right] &= E\left[\left\{Y_t - F\left(X_t\beta\right)\right\}^2\right] \\ &= Pr\left(Y_t = 1\right)\left[1 - F\left(X_t\beta\right)\right]^2 + Pr\left(Y_t = 0\right)\left[0 - F\left(X_t\beta\right)\right]^2 \\ &= F\left(X_t\beta\right)\left[1 - F\left(X_t\beta\right)\right]^2 + \left[1 - F\left(X_t\beta\right)\right]F\left(X_t\beta\right)^2 \\ &= F\left(X_t\beta\right)\left[1 - F\left(X_t\beta\right)\right] \end{aligned}$$

As *ML* is equivalent to weighted nonlinear least squares for binary response models, the asymptotic covariance matrix for $n^{1/2}\left(\hat{\beta} - \beta\right)$ is $\left(n^{-1}X^T\Psi X\right)^{-1}$ where Ψ is a diagonal matrix with elements $\frac{f(X_t\beta)^2}{F(X_t\beta)[1-F(X_t\beta)]}$. In the logit case, Ψ simplifies to λ (see Davidson and MacKinnon, p. 517-518).

4.3.2 Estimated asymptotic covariance for MLE of $\hat{\theta}$

There are (at least) three common estimators for the variance of $\hat{\theta}_{MLE}$:[7]
(i) $\left[-H\left(\hat{\theta}\right)\right]^{-1}$ the negative inverse of Hessian evaluated at $\hat{\theta}_{MLE}$,
(ii) $\left[g\left(\hat{\theta}\right)g\left(\hat{\theta}\right)^T\right]^{-1}$ the outer product of gradient (*OPG*) or Berndt, Hall, Hall, and Hausman (*BHHH*) estimator,

[7] This section draws heavily from Davidson and MacKinnon [1993], pp. 260-267.

(iii) $\left[\Im\left(\hat{\theta}\right)\right]^{-1}$ inverse of information matrix or negative expected value of Hessian, where the following definitions apply:

- *MLE* is defined as the solution to the first order conditions (*FOC*): $g\left(Y,\hat{\theta}\right) = 0$ where gradient or score vector g is defined by $g^T(Y,\theta) = D_\theta \ell(Y,\theta)$ (since $D_\theta \ell$ is row vector, g is column vector of partial derivatives of with respect to θ).

- Define $G(g,\theta)$ as the matrix of contributions to the gradient (*CG* matrix) with typical element $G_{ti}(g,\theta) \equiv \frac{\partial \ell_t(Y,\theta)}{\partial \theta_i}$.

- $H(Y,\theta)$ is the Hessian matrix for the log-likelihood with typical element $H_{ij}(Y,\theta) \equiv \frac{\partial^2 \ell_t(Y,\theta)}{\partial \theta_i \partial \theta_j}$.

- Define the expected average Hessian for sample of size n as $H_n(\theta) \equiv E_\theta\left[n^{-1}H(Y,\theta)\right]$.

- The limiting Hessian or asymptotic Hessian (if it exists) is $H(\theta) \equiv \lim_{n\to\infty} H_n(\theta)$ (the matrix is negative semidefinite).

- Define the information in observation t as $\Im_t(\theta)$ a $k \times k$ matrix with typical element $(\Im_t(\theta))_{ij} \equiv E_\theta\left[G_{ti}(\theta) G_{tj}(\theta)\right]$ (the information matrix is positive semidefinite).

- The average information matrix is $\Im_n(\theta) \equiv n^{-1}\sum_{t=1}^{n} \Im_t(\theta) = n^{-1}\Im_n$ and the limiting information matrix or asymptotic information matrix (if it exists) is $\Im(\theta) \equiv \lim_{n\to\infty} \Im_n(\theta)$.

The short explanation for these variance estimators is that *ML* estimators (under suitable regularity conditions) achieve the Cramer-Rao lower bound for consistent

estimators.[8] That is,

$$Asy.Var\left[\hat{\theta}\right] = \left\{-E\left[\frac{\partial^2 \ell\left(Y,\theta\right)}{\partial\theta\partial\theta^T}\right]\right\}^{-1} = \left\{E\left[\left(\frac{\partial \ell\left(Y,\theta\right)}{\partial\theta}\right)\left(\frac{\partial \ell\left(Y,\theta\right)}{\partial\theta^T}\right)\right]\right\}^{-1}$$

The expected outer product of the gradient (*OPG*) is an estimator of the inverse of the variance matrix for the gradient. Roughly speaking, the inverse of the gradient function yields *MLE* (type 2) parameter estimates and the inverse of expected *OPG* estimates the parameter variance matrix (see Berndt, Hall, Hall, and Hausman [1974]). Also, the expected value of the Hessian equals the negative of the information matrix.[9] In turn, the inverse of the information matrix is an estimator for the estimated parameter variance matrix.

Example: Consider the *MLE* of a standard linear regression model with *DGP*: $Y = X\beta + \varepsilon$ where $\varepsilon \sim N\left(0, \sigma^2 I\right)$ and $E\left[X^T\varepsilon\right] = 0$. Of course, the *MLE* for β is $b = \left(X^T X\right)^{-1} X^T Y$ as

$$g\left(\beta\right) \equiv \frac{\partial \ell\left(Y,\beta\right)}{\partial\beta} = -\frac{1}{\sigma^2}\begin{bmatrix} X_1^T\left(Y - X\beta\right) \\ \vdots \\ X_p^T\left(Y - X\beta\right) \end{bmatrix}$$

---

[8] See Theil [1971], pp. 384-385 and Amemiya [1985], pp. 14-17.

$$E\left[\frac{\partial^2 \ell}{\partial\theta\partial\theta^T}\right] = E\left[\frac{\partial}{\partial\theta}\left(\frac{1}{L}\frac{\partial L}{\partial\theta^T}\right)\right]$$

by the chain rule

$$\begin{aligned} &= E\left[-\frac{1}{L^2}\frac{\partial L}{\partial\theta}\frac{\partial L}{\partial\theta^T} + \frac{1}{L}\frac{\partial^2 L}{\partial\theta\partial\theta^T}\right] \\ &= E\left[-\left(\frac{1}{L}\frac{\partial L}{\partial\theta}\right)\left(\frac{\partial L}{\partial\theta^T}\frac{1}{L}\right)\right] + \int \frac{1}{L}\frac{\partial^2 L}{\partial\theta\partial\theta^T} L dx \\ &= E\left[-\left(\frac{1}{L}\frac{\partial L}{\partial\theta}\right)\left(\frac{\partial L}{\partial\theta^T}\frac{1}{L}\right)\right] + \int \frac{\partial^2 L}{\partial\theta\partial\theta^T} dx \\ &= -E\left[\frac{\partial \ell}{\partial\theta}\frac{\partial \ell}{\partial\theta^T}\right] + \int \frac{\partial^2 L}{\partial\theta\partial\theta^T} dx \end{aligned}$$

since the regulatory conditions essentially make the order of integration and differentiation interchangeable the last term can be rewritten

$$\int \frac{\partial^2 L}{\partial\theta\partial\theta^T} dx = \frac{\partial}{\partial\theta}\int \frac{\partial L}{\partial\theta^T} dx = \frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta^T}\int L dx = 0$$

Now we have

$$E\left[\frac{\partial^2 \ell}{\partial\theta\partial\theta^T}\right] = -E\left[\frac{\partial \ell}{\partial\theta}\frac{\partial \ell}{\partial\theta^T}\right]$$

[9] This is motivated by the fact that $plim\frac{1}{n}\sum_{i=1}^{n} g(y_i) = E[g(y)]$ for a random sample provided the first two moments of $g(y)$ are finite (see Greene [1997], ch. 4).

where X_j refers to column j of X. Substituting $X\beta + \varepsilon$ for Y produces

$$\left(\frac{\partial \ell(Y,\beta)}{\partial \beta}\right)\left(\frac{\partial \ell(Y,\beta)}{\partial \beta^T}\right) = \left(\frac{1}{\sigma^2}\right)^2 \begin{bmatrix} X_1^T\varepsilon\varepsilon^T X_1 & \cdots & X_1^T\varepsilon\varepsilon^T X_p \\ \vdots & \ddots & \vdots \\ X_p^T\varepsilon\varepsilon^T X_1 & \cdots & X_p^T\varepsilon\varepsilon^T X_p \end{bmatrix}$$

Now,

$$E\left[\left(\frac{\partial \ell(Y,\beta)}{\partial \beta}\right)\left(\frac{\partial \ell(Y,\beta)}{\partial \beta^T}\right)\right] = \left(\frac{1}{\sigma^2}\right) \begin{bmatrix} X_1^T X_1 & \cdots & X_1^T X_p \\ \vdots & \ddots & \vdots \\ X_p^T X_1 & \cdots & X_p^T X_p \end{bmatrix}$$

Since

$$H(\beta) \equiv \frac{\partial^2 \ell(Y,\beta)}{\partial\beta\partial\beta^T} = -\left(\frac{1}{\sigma^2}\right) \begin{bmatrix} X_1^T X_1 & \cdots & X_1^T X_p \\ \vdots & \ddots & \vdots \\ X_p^T X_1 & \cdots & X_p^T X_p \end{bmatrix}$$

we have

$$E\left[\left(\frac{\partial \ell(Y,\beta)}{\partial \beta}\right)\left(\frac{\partial \ell(Y,\beta)}{\partial \beta^T}\right)\right] = -E\left[\frac{\partial^2 \ell(Y,\beta)}{\partial\beta\partial\beta^T}\right]$$

and the demonstration is complete as

$$\begin{aligned} Asy.Var[b] &= \left\{E\left[\left(\frac{\partial \ell(Y,\beta)}{\partial \beta}\right)\left(\frac{\partial \ell(Y,\beta)}{\partial \beta^T}\right)\right]\right\}^{-1} \\ &= -\left\{E\left[\frac{\partial^2 \ell(Y,\beta)}{\partial\beta\partial\beta^T}\right]\right\}^{-1} \\ &= \sigma^2\left(X^T X\right)^{-1} \end{aligned}$$

A more complete explanation (utilizing results and notation developed in the appendix) starts with the *MLE* first order condition (*FOC*) $g\left(\hat{\theta}\right) = 0$. Now, a Taylor series expansion of the likelihood *FOC* around θ yields $0 = g\left(\hat{\theta}\right) \approx g(\theta) + H\left(\bar{\theta}\right)\left(\hat{\theta} - \theta\right)$ where $\bar{\theta}$ is convex combination (perhaps different for each row) of θ and $\hat{\theta}$. Solve for $\left(\hat{\theta} - \theta\right)$ and rewrite so every term is $O(1)$

$$n^{1/2}\left(\hat{\theta} - \theta\right) = -\left[n^{-1}H\left(\bar{\theta}\right)\right]^{-1}\left[n^{-1/2}g(\theta)\right]$$

By *WULLN* (weak uniform law of large numbers), the first term is asymptotically nonstochastic, by *CLT* (the central limit theorem) the second term is asymptotically normal, so $n^{1/2}\left(\hat{\theta} - \theta\right)$ is asymptotically normal. Hence, the asymptotic variance of $n^{1/2}\left(\hat{\theta} - \theta\right)$ is the asymptotic expectation of $n\left(\hat{\theta} - \theta\right)\left(\hat{\theta} - \theta\right)^T$.

Since $n^{1/2}\left(\hat{\theta}-\theta\right) \stackrel{a}{=} -\left[n^{-1}H\left(\theta\right)\right]^{-1}\left[n^{-1/2}g\left(\theta\right)\right]$, the asymptotic variance is $\left(-H^{-1}\left(\theta\right)\right)\left(n^{-1}E_{\theta}\left[g\left(\theta\right)g^{T}\left(\theta\right)\right]\right)\left[-H^{-1}\left(\theta\right)\right]$. Simplifying yields

$$Asym.Var\left[n^{1/2}\left(\hat{\theta}-\theta\right)\right] = H^{-1}\left(\theta\right)\Im\left(\theta\right)H^{-1}\left(\theta\right)$$

This can be simplified since $H\left(\theta\right) = -\Im\left(\theta\right)$ by *LLN*. Hence,

$$Asy.Var\left[n^{1/2}\left(\hat{\theta}-\theta\right)\right] = -H^{-1}\left(\theta\right) = \Im^{-1}\left(\theta\right)$$

And the statistic relies on estimation of $H^{-1}\left(\theta\right)$ or $\Im^{-1}\left(\theta\right)$.

- A common estimator of the empirical Hessian is

  $$\hat{H} \equiv H_{n}\left(Y,\hat{\theta}\right) = n^{-1}D_{\theta\theta}^{2}\ell_{t}\left(Y,\hat{\theta}\right)$$

  (*LLN* and consistency of $\hat{\theta}$ guarantee consistency of $\hat{H}$ for $H\left(\theta\right)$).

- The *OPG* or *BHHH* estimator is

  $$\Im_{OPG} \equiv n^{-1}\sum_{t=1}^{n}D_{\theta}^{T}\ell_{t}\left(Y,\hat{\theta}\right)D_{\theta}\ell_{t}\left(Y,\hat{\theta}\right) = n^{-1}G^{T}\left(\hat{\theta}\right)G\left(\hat{\theta}\right)$$

  (consistency is guaranteed by *CLT* and *LLN* for the sum).

- The third estimator evaluates the expected values of the second derivatives of the log-likelihood at $\hat{\theta}$. Since this form is not always known, this estimator may not be available. However, as this estimator does not depend on the realization of Y it is less noisy than the other estimators.

We round out this discussion of *MLE* by reviewing a surprising case where *MLE* is not the most efficient estimator. Next, we discuss James-Stein shrinkage estimators.

4.4 James-Stein shrinkage estimators

Stein [1955] showed that when estimating K parameters from independent normal observations with (for simplicity) unit variance, we can uniformly improve on the conventional maximum likelihood estimator in terms of expected squared error loss for $K > 2$. James and Stein [1961] determined such a shrinkage estimator can be written as a function of the maximum likelihood estimator $\widehat{\theta}$

$$\theta^{*} = \widehat{\theta}\left(1-\frac{a}{\widehat{\theta}^{T}\widehat{\theta}}\right)$$

where $0 \leq a \leq 2(K-2)$. The expected squared error loss of the James-Stein estimator θ^* is

$$\begin{aligned}
\rho(\theta, \theta^*) &= E\left[(\theta - \theta^*)^T (\theta - \theta^*)\right] \\
&= E\left[\left\{\left(\widehat{\theta} - \theta\right) - \frac{a\widehat{\theta}}{\widehat{\theta}^T\widehat{\theta}}\right\}^T \left\{\left(\widehat{\theta} - \theta\right) - \frac{a\widehat{\theta}}{\widehat{\theta}^T\widehat{\theta}}\right\}\right] \\
&= E\left[\left(\widehat{\theta} - \theta\right)^T \left(\widehat{\theta} - \theta\right)\right] - 2aE\left[\left(\widehat{\theta} - \theta\right)^T \frac{\widehat{\theta}}{\widehat{\theta}^T\widehat{\theta}}\right] \\
&\quad + a^2 E\left[\frac{\widehat{\theta}^T\widehat{\theta}}{\left(\widehat{\theta}^T\widehat{\theta}\right)^2}\right] \\
&= E\left[\left(\widehat{\theta} - \theta\right)^T \left(\widehat{\theta} - \theta\right)\right] - 2aE\left[\frac{\widehat{\theta}^T\widehat{\theta}}{\widehat{\theta}^T\widehat{\theta}}\right] + 2a\theta^T E\left[\frac{\widehat{\theta}}{\widehat{\theta}^T\widehat{\theta}}\right] \\
&\quad + a^2 E\left[\frac{\widehat{\theta}^T\widehat{\theta}}{\left(\widehat{\theta}^T\widehat{\theta}\right)^2}\right] \\
&= E\left[\left(\widehat{\theta} - \theta\right)^T \left(\widehat{\theta} - \theta\right)\right] - 2a + 2a\theta^T E\left[\frac{\widehat{\theta}}{\widehat{\theta}^T\widehat{\theta}}\right] + a^2 E\left[\frac{1}{\widehat{\theta}^T\widehat{\theta}}\right]
\end{aligned}$$

This can be further simplified by exploiting the following theorems; we conclude this section with Judge and Bock's [1978, p. 322-3] proof following discussion of the James-Stein shrinkage estimator.

Theorem 4.1 $E\left[\frac{\widehat{\theta}}{\widehat{\theta}^T\widehat{\theta}}\right] = \theta E\left[\frac{1}{\chi^2_{(K+2,\lambda)}}\right]$ *where* $\widehat{\theta}^T\widehat{\theta} \sim \chi^2_{(K,\lambda)}$ *and* $\lambda = \theta^T\theta$ *is a noncentrality parameter*.[10]

Using

$$\begin{aligned}
E\left[\left(\widehat{\theta} - \theta\right)^T \left(\widehat{\theta} - \theta\right)\right] &= E\left[\widehat{\theta}^T\widehat{\theta}\right] - 2\theta^T E\left[\widehat{\theta}\right] + \theta^T\theta \\
&= K + \lambda - 2\lambda + \lambda = K
\end{aligned}$$

for the first term, a convenient substitution for one in the second term, and the above theorem for the third term, we rewrite the squared error loss (from above)

$$\rho(\theta, \theta^*) = E\left[\left(\widehat{\theta} - \theta\right)^T \left(\widehat{\theta} - \theta\right)\right] - 2a + 2a\theta^T E\left[\frac{\widehat{\theta}}{\widehat{\theta}^T\widehat{\theta}}\right] + a^2 E\left[\frac{1}{\widehat{\theta}^T\widehat{\theta}}\right]$$

[10] We adopt the convention the noncentrality parameter is the sum of squared means $\theta^T\theta$; others, including Judge and Bock [1978], employ $\frac{\theta^T\theta}{2}$.

as

$$\rho(\theta,\theta^*) = K - 2aE\left[\frac{\chi^2_{(K-2,\lambda)}}{\chi^2_{(K-2,\lambda)}}\right] + 2a\theta^T\theta E\left[\frac{1}{\chi^2_{(K+2,\lambda)}}\right] + a^2 E\left[\frac{1}{\chi^2_{(K,\lambda)}}\right]$$

Theorem 4.2 *For any real-valued function f and positive definite matrix A,*

$$\begin{aligned} E\left[f\left(\widehat{\theta}^T\widehat{\theta}\right)\left(\widehat{\theta}^T A\widehat{\theta}\right)\right] &= E\left[f\left(\chi^2_{(K+2,\lambda)}\right) tr(A)\right] \\ &\quad + E\left[f\left(\chi^2_{(K+4,\lambda)}\right)\right]\left(\theta^T A\theta\right) \end{aligned}$$

where $tr(A)$ is trace of A.

Letting $f\left(\widehat{\theta}^T\widehat{\theta}\right) = \frac{1}{\chi^2_{(K-2,\lambda)}}$ and $A = I$ with rank $K-2$,

$$-2aE\left[\frac{\chi^2_{(K-2,\lambda)}}{\chi^2_{(K-2,\lambda)}}\right] = -2aE\left[\frac{K-2}{\chi^2_{(K,\lambda)}}\right] - 2a\theta^T\theta E\left[\frac{1}{\chi^2_{(K+2,\lambda)}}\right]$$

and

$$\rho(\theta,\theta^*) = K - a\left[2(K-2) - a\right]E\left[\frac{1}{\chi^2_{(K,\lambda)}}\right]$$

Hence, $\rho(\theta,\theta^*) = K - a\left[2(K-2)-a\right]E\left[\frac{1}{\chi^2_{(K,\lambda)}}\right] \leq \rho\left(\theta,\widehat{\theta}\right) = K$ for all θ if $0 < a < 2(K-2)$ with strict inequality for some $\theta^T\theta$.

Now, we can find the optimal James-Stein shrinkage estimator. Solving the first order condition

$$\begin{aligned} \frac{\partial\rho(\theta,\theta^*)}{\partial a} &= 0 \\ (-2(K-2) - a + 2a)E\left[\frac{1}{\chi^2_{(K,\lambda)}}\right] &= 0 \end{aligned}$$

leads to $a^* = K-2$; hence, $\theta^* = \widehat{\theta}\left(1 - \frac{K-2}{\widehat{\theta}^T\widehat{\theta}}\right)$. As $E\left[\frac{1}{\chi^2_{(K,\lambda)}}\right] = \frac{1}{K-2}$, the James-Stein estimator has minimum expected squared error loss when $\theta = 0$,

$$\begin{aligned} \rho(\theta,\theta^*) &= K - (K-2)^2 E\left[\frac{1}{\chi^2_{(K,\lambda)}}\right] \\ &= K - (K-2) = 2 \end{aligned}$$

and its *MSE* approaches that for the *MLE* as $\lambda = \theta^T\theta$ approaches infinity. Next, we sketch proofs of the theorems.

Stein [1966] identified a key idea used in the proofs. Suppose a $J \times 1$ random vector w is distributed as $N(\theta, I)$, then its quadratic form $w^T w$ has a noncentral

$\chi^2_{(J,\lambda)}$ where $\lambda = \theta^T\theta$. This quadratic form can be regarded as having a central $\chi^2_{(J+2H)}$ where H is a Poisson random variable with parameter $\frac{\lambda}{2}$. Hence,

$$\begin{aligned} E\left[f\left(\chi^2_{(J,\lambda)}\right)\right] &= E_H\left[HE\left[f\left(\chi^2_{(J+2H)}\right)\right]\right] \\ &= \sum_{t=0}^{\infty}\left(\frac{\lambda}{2}\right)^t \frac{\exp\left[-\frac{\lambda}{2}\right]}{t!} E\left[f\left(\chi^2_{(J+2t)}\right)\right] \end{aligned}$$

Now, we proceed with proofs to the above theorems.

Theorem 4.3 $E\left[\frac{\widehat{\theta}}{\widehat{\theta}^T\widehat{\theta}}\right] = \theta E\left[\frac{1}{\chi^2_{(K+2,\lambda)}}\right]$.

Proof. Write

$$\begin{aligned} E\left[f\left(w^2\right)w\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f\left(w^2\right) w \exp\left[-\frac{(w-\theta)^2}{2}\right] dw \\ &= \exp\left[-\frac{\theta^2}{2}\right]\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f\left(w^2\right) w \exp\left[-\frac{w^2}{2}+\theta w\right] dw \end{aligned}$$

Rewrite as

$$\exp\left[-\frac{\theta^2}{2}\right]\frac{1}{\sqrt{2\pi}}\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f\left(w^2\right)\exp\left[-\frac{w^2}{2}+\theta w\right] dw$$

complete the square and write

$$\exp\left[-\frac{\theta^2}{2}\right]\frac{\partial}{\partial\theta}\left\{E\left[f\left(w^2\right)\right]\exp\left[\frac{\theta^2}{2}\right]\right\}$$

Since $w \sim N(\theta,1)$, $w^2 \sim \chi^2_{(1,\theta^2)}$. Now, apply Stein's observation

$$\begin{aligned} & \exp\left[-\frac{\theta^2}{2}\right]\frac{\partial}{\partial\theta}\left\{E\left[f\left(w^2\right)\right]\exp\left[\frac{\theta^2}{2}\right]\right\} \\ = & \exp\left[-\frac{\theta^2}{2}\right]\frac{\partial}{\partial\theta}\left\{\exp\left[\frac{\theta^2}{2}\right]\sum_{j=0}^{\infty}\left(\frac{\theta^2}{2}\right)^j\frac{\exp\left[-\frac{\theta^2}{2}\right]}{j!}E\left[f\left(\chi^2_{(1+2j)}\right)\right]\right\} \\ = & \exp\left[-\frac{\theta^2}{2}\right]\frac{\partial}{\partial\theta}\left\{\sum_{j=0}^{\infty}\left(\frac{\theta^2}{2}\right)^j\frac{1}{j!}E\left[f\left(\chi^2_{(1+2j)}\right)\right]\right\} \end{aligned}$$

Taking the partial derivative yields

$$\exp\left[-\frac{\theta^2}{2}\right]\theta\left\{\sum_{j=1}^{\infty}\left(\frac{\theta^2}{2}\right)^{j-1}\frac{1}{(j-1)!}E\left[f\left(\chi^2_{(3+2(j-1))}\right)\right]\right\}$$

or

$$E\left[f\left(w^2\right)w\right] = \theta E\left[f\left(\chi^2_{(3,\theta^2)}\right)\right]$$

For the multivariate case at hand, this implies

$$E\left[\frac{\widehat{\theta}}{\widehat{\theta}^T\widehat{\theta}}\right] = \theta E\left[\frac{1}{\chi^2_{(K+2,\lambda)}}\right]$$

■

Lemma 4.1 $E\left[f\left(w^2\right)w^2\right] = E\left[f\left(\chi^2_{(3,\theta^2)}\right)\right] + \theta^2 E\left[f\left(\chi^2_{(5,\theta^2)}\right)\right]$.

Proof. Let $z \sim \chi^2_{(1,\theta^2)}$.

$$\begin{aligned} E\left[f\left(w^2\right)w^2\right] &= E\left[f\left(z\right)z\right] \\ &= \sum_{j=0}^{\infty}\left(\frac{\theta^2}{2}\right)^j \frac{1}{j!}\exp\left[-\frac{\theta^2}{2}\right] E\left[f\left(\chi^2_{(1+2j)}\right)\right] \end{aligned}$$

Since

$$E\left[f\left(\chi^2_{(n)}\right)\chi^2_{(n)}\right] = \int_0^{\infty} \frac{f\left(s\right)s\exp\left[-\frac{s}{2}\right]s^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)2^{\frac{n}{2}}}ds$$

combining terms involving powers of s and rewriting

$$\Gamma\left(t\right) = \int_0^{\infty} x^{t-1}\exp\left[-x\right]dx$$

for $t > 0$, as $\Gamma\left(t=1\right) = t\Gamma\left(t\right)$, leads to

$$\begin{aligned} & n\int_0^{\infty} \frac{f\left(s\right)\exp\left[-\frac{s}{2}\right]s^{\frac{n+2}{2}-1}}{\Gamma\left(\frac{n+2}{2}\right)2^{\frac{n+2}{2}}}ds \\ = \;& nE\left[f\left(\chi^2_{(n+2)}\right)\right] \end{aligned}$$

Now,

$$\begin{aligned} & \sum_{j=0}^{\infty}\left(\frac{\theta^2}{2}\right)^j \frac{1}{j!}\exp\left[-\frac{\theta^2}{2}\right] E\left[f\left(\chi^2_{(1+2j)}\right)\right] \\ = \;& \sum_{j=0}^{\infty}\left(\frac{\theta^2}{2}\right)^j \frac{1}{j!}\exp\left[-\frac{\theta^2}{2}\right] \left(1+2j\right) E\left[f\left(\chi^2_{(3+2j)}\right)\right] \end{aligned}$$

Rewrite as

$$\begin{aligned} & \sum_{j=0}^{\infty}\left(\frac{\theta^2}{2}\right)^j \frac{1}{j!}\exp\left[-\frac{\theta^2}{2}\right] E\left[f\left(\chi^2_{(3+2j)}\right)\right] \\ & +2\frac{\theta^2}{2}\sum_{j=0}^{\infty}\left(\frac{\theta^2}{2}\right)^{j-1} \frac{1}{(j-1)!}\exp\left[-\frac{\theta^2}{2}\right] E\left[f\left(\chi^2_{(5+2(j-1))}\right)\right] \end{aligned}$$

Again, apply Stein's observation to produce

$$E\left[f\left(w^2\right)w^2\right] = E\left[f\left(\chi^2_{(3,\theta^2)}\right)\right] + \theta^2 E\left[f\left(\chi^2_{(5,\theta^2)}\right)\right]$$

■

Theorem 4.4

$$\begin{aligned} E\left[f\left(\widehat{\theta}^T\widehat{\theta}\right)\left(\widehat{\theta}^T A\widehat{\theta}\right)\right] &= E\left[f\left(\chi^2_{(K+2,\lambda)}\right)tr\left(A\right)\right] \\ &\quad + E\left[f\left(\chi^2_{(K+4,\lambda)}\right)\right]\left(\theta^T A\theta\right) \end{aligned}$$

Proof. Let P be an orthogonal matrix such that $PAP^T = D$, a diagonal matrix with eigenvalues of A, $d_j > 0$, along the diagonal. Define vector $\omega = Pw \sim N\left(P\theta, I\right)$. Since

$$\omega^T\omega = w^T P^T P w = w^T w$$

and

$$\omega^T D\omega = \omega^T P^T AP\omega = w^T Aw$$

$$E\left[f\left(\omega^T\omega\right)\omega^T D\omega\right] = \sum_{i=1}^{J} d_i E\left[E\left[f\left(\omega_i^2 + \sum_{j\neq i}\omega_j^2\right)\omega_i^2 \mid \omega_j, i \neq j\right]\right]$$

Using the lemma, this can be expressed as

$$\sum_{i=1}^{J} d_i \left\{ \begin{array}{c} E\left[f\left(\chi^2_{\left(3,\left(p_i^T\theta\right)^2\right)} + \sum_{j\neq i}\omega_j^2\right)\right] \\ + \left(p_i^T\theta\right)^2 E\left[f\left(\chi^2_{\left(5,\left(p_i^T\theta\right)^2\right)} + \sum_{j\neq i}\omega_j^2\right)\right] \end{array} \right\}$$

where p_i^T is the ith row of P. Since $\sum_{i=1}^{J} d_i\left(p_i^T\theta\right)^2 = \theta^T A\theta$ and $\sum_{i=1}^{J} d_i = tr\left(A\right)$,

$$\begin{aligned} E\left[f\left(\widehat{\theta}^T\widehat{\theta}\right)\left(\widehat{\theta}^T A\widehat{\theta}\right)\right] &= E\left[f\left(\chi^2_{(K+2,\lambda)}\right)tr\left(A\right)\right] \\ &\quad + E\left[f\left(\chi^2_{(K+4,\lambda)}\right)\right]\left(\theta^T A\theta\right) \end{aligned}$$

■

4.5 Summary

This chapter has briefly reviewed loss functions, nonlinear regression, maximum likelihood estimation, and some alternative estimation methods (including James-Stein shrinkage estimators). It is instructive to revisit nonlinear regression (especially, *GNR*) in the next chapter when we address specification and estimation of discrete choice models.

4.6 Additional reading

Poirier [1995] provides a nice discussion of loss functions. Conditional linear loss functions lead to quantile regression (see Koenker and Bassett [1978], Koenker [2005], and Koenker [2009] for an **R** computational package). Shugan and Mitra [2008] offer an intriguing discussion of when and why non-averaging statistics (e.g., maximum and variance) explain more variance than averaging metrics. Maximum likelihood estimation is discussed by a broad range of authors including Davidson and MacKinnon [1993], Greene [1997], Amemiya [1985], Rao [1973], and Theil [1971]. Stigler [2007] provides a fascinating account of the history of maximum likelihood estimation including the pioneering contributions of Gauss and Fisher as well as their detractors. The nonlinear regression section draws heavily from a favorite reference, Davidson and MacKinnon [2003]. Their chapter 6 and references therein provide a wealth of ideas related to estimation and specification of nonlinear models.