Problem 1. Define $H \coloneqq X(X^\top X)^{-1}X^\top$, where X is a non-stochastic $n \times p$ full rank matrix with $p \leq n$. Show that

1. H is idempotent and symmetric, meaning that $H^2 = H$ and $H^\top = H$.
2. the eigenvalues of H are either 0 or 1.
3. H is a projection matrix onto the column space of X, $\mathcal{S}(X)$. Is this still the case if the columns of X are not linearly independent?
4. the trace of H, $\text{tr}(H)$, is equal to p and thus $\text{rank}(H) = p$.
5. $X^\top X$ is invertible.

Solution

1. symmetry is trivial. For idempotency, $H_X H_X = X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = XI_p(X^\top X)^{-1}X^\top = H_X$.
2. If v is an eigenvector of H associated to the eigenvalue λ, then $Hv = \lambda v$ by definition. But H is idempotent, so $H^2 v = \lambda Hv = \lambda^2 v$ and the only solutions of $\lambda^2 = \lambda$ are $\{0, 1\}$.
3. The matrix H is symmetric and idempotent. It remains to show its image is $\mathcal{S}(X)$. For any $y \in \mathbb{R}^n$, $Hy = X\widehat{\beta}$ with $\widehat{\beta} = (X^\top X)^{-1}X^\top y \in \mathbb{R}^p$. Thus $\text{im}(H) \subseteq \mathcal{S}(X)$, while at the same time $HX = X$, so $\text{im}(H) \supseteq \mathcal{S}(X)$. H is not well-defined if X does not have rank p since the inverse $X^\top X$ does not exist.
4. The trace is invariant to cyclic permutations of its arguments, so

$$\text{tr}(H) = \text{tr}\left(X^\top X(X^\top X)^{-1}\right) = \text{tr}(I_p) = p.$$

   The trace is also equal to the sum of the eigenvalues of H, which are either 0 or 1. There must be p non-zero eigenvalues, so by the spectral theorem $\text{rank}(H) = p$.
5. The matrix $X^\top X : \mathbb{R}^p \to \mathbb{R}^p$ is invertible if and only if it is bijective. Assume $X^\top Xv = 0_p$ for some $v \in \mathbb{R}^p$. Then,

$$\|Xv\|^2 = v^\top X^\top Xv = v^\top 0_p = 0.$$

   It thus suffices to show that X is injective. Note that X has p linearly independent columns, so $\text{rank}(X) \coloneqq \dim(\text{im}(X)) = p$ and $X : \mathbb{R}^p \to \mathbb{R}^n$, so by the rank-nullity theorem, $\dim(\ker\{X\}) + \text{rank}(X) = \dim(\mathbb{R}^p)$ and $\dim(\ker\{X\}) = 0$. Since the kernel of X is trivial, X is injective and so $v = 0_p$, proving that $X^\top X$ is one-to-one. The mapping $X^\top X$ is surjective because of the rank-nullity theorem; again $X^\top X : \mathbb{R}^p \to \mathbb{R}^p$ so $\dim(\ker\{X^\top X\}) + \text{rank}(X^\top X) = \dim(\mathbb{R}^p)$ and $\dim(\ker\{X^\top X\}) = 0$ by the previous part, so $\text{rank}(X^\top X) = p$ and this completes the proof.

   If the columns of X are linearly dependent, there exists a non-zero vector $v \in \mathbb{R}^p$ such that $Xv = 0_p$, so $X^\top Xv = 0_p$ and $X^\top X$ is not injective, thus not invertible.

Problem 2. Show that orthogonal projection matrices[1] are unique: if P and Q are orthogonal projection matrices onto a subspace $\mathcal{V}$ of $\mathbb{R}^n$, then $P = Q$.

Solution

There are many ways to prove this. First, the column vectors of P are elements of $\mathcal{V}$. Consider a basis V of p orthogonal vectors in $\mathcal{V}$ and a basis of $n - p$ vectors W for $\mathcal{V}^\perp$. We can express the ith column vector of P as $p_i = V\alpha + W\gamma$ for some coefficients $\alpha \in \mathbb{R}^p, \gamma \in \mathbb{R}^{n-p}$. Because P is idempotent, $Pp_1 = p_1$ and so $\gamma = 0_{n-p}$. This shows columns of $P \in \mathcal{V}$, so $QP = P$ since Q is a projector. Similarly, $PQ = Q$. Using symmetry,

$$Q = PQ = P^\top Q^\top = (QP)^\top = P^\top = P.$$

[1] Note: the projection is orthogonal, not the matrix — the later is not invertible if $p < n$! The three defining properties of an orthogonal projection matrices on to $\mathcal{V}$ are (1) $Pv = v$ for any $v \in \mathcal{S}(\mathcal{V})$, (2) symmetry and (3) idempotency.

Compiled 09/12/2019 at 18:51

Alternatively: for any $v \in \mathcal{V}$, $v = P\beta$ for some β. Pre-multiply both sides by P and use the idempotency of projection matrices to get $Pv = PP\beta = P\beta = v$.

We thus have $Pv = v = Qv$ for any $v \in \mathcal{V}$. Since any vector $x \in \mathbb{R}^n$ can be uniquely decomposed into two orthogonal vectors $x = v + w$, where $v \in \mathcal{V}$ and $w \in \mathcal{V}^\perp$, $Qx = v = Px$ for any $x \in \mathbb{R}^n$ and thus $P = Q$.

Problem 3. Suppose the $n \times p$ full-rank design matrix X can be written as $[X_1\ X_2]$ with blocks X_1, an $n \times p_1$ matrix, and X_2, an $n \times p_2$ matrix. Show that $H - H_1$ is an orthogonal projection matrix. ($H_1 = X_1(X_1^\top X_1)^{-1}X_1^\top$)

Solution

The key is to note that $HX_1 = X_1$ since the columns of X_1 are in $\mathcal{S}(X)$. It follows that $HH_1 = H_1$ and, by transposing, that $H_1H = H_1$. The matrix $H - H_1$ is symmetric since both H and H_1 are symmetric. The idempotency follows from the observation that

$$\begin{aligned}(H - H_1)(H - H_1) &= HH - H_1H - HH_1 + H_1H_1 \\ &= H - H_1 \pm H_1H_1 \\ &= H - H_1.\end{aligned}$$

Problem 4. Suppose that $A, X \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$. Show that

1. $\frac{\partial}{\partial x}Ax = A^\top$;
2. $\frac{\partial}{\partial x}x^\top Ax = (A + A^\top)x$; $\quad$ [Note the special case $\frac{\partial}{\partial x}x^\top x = 2x$.]
3. $\frac{\partial}{\partial X}\mathrm{tr}(X) = I_n$.

Solution

a) Denote $y = Ax$. Hence, $y_i = \sum_{j=1}^n A_{ij}x_j$ and thus $\frac{\partial}{\partial x_j}y_i = A_{ij}$. We obtain

$$\frac{\partial}{\partial x}y = \begin{bmatrix} \frac{\partial}{\partial x_1}y_1 & \frac{\partial}{\partial x_1}y_2 & \cdots & \frac{\partial}{\partial x_1}y_n \\ \frac{\partial}{\partial x_2}y_1 & \frac{\partial}{\partial x_2}y_2 & \cdots & \frac{\partial}{\partial x_2}y_n \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_n}y_1 & \frac{\partial}{\partial x_n}y_2 & \cdots & \frac{\partial}{\partial x_n}y_n \end{bmatrix} = \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix} = A^\top.$$

b) Denote $y = x^\top Ax = \sum_{i=1}^n \sum_{j=1}^n A_{ij}x_ix_j$. We have

$$\begin{aligned}\frac{\partial}{\partial x_k}y &= \sum_{i\neq k} A_{ki}x_i + \sum_{i\neq k} A_{ik}x_i + 2A_{kk}x_k \\ &= \sum_{i=1}^n A_{ki}x_i + \sum_{i=1}^n A_{ik}x_i = (Ax)_k + (A^\top x)_k = (Ax + A^\top x)_k.\end{aligned}$$

Thus

$$\frac{\partial}{\partial x}y = Ax + A^\top x = (A + A^\top)x.$$

c) Denote $y = \mathrm{tr}(X) = \sum_{i=1}^n X_{ii}$. Then $\frac{\partial}{\partial X_{ij}}y = \delta_{ij}$, where

$$\delta_{ij} = \begin{cases} 1, & \text{si } i = j, \\ 0, & \text{si } i \neq j \end{cases} \tag{1}$$

is the Kronecker delta. Thus $\frac{\partial}{\partial X}y = I_n$.

Problem 5. Let X be an $n \times p$ full rank real matrix with $p \leq n$ and Ω an $n \times n$ positive definite matrix, meaning that $v^\top \Omega v > 0$ for all $v \in \mathbb{R}^n \setminus \{0_n\}$.

1. Show that $B = X^\top \Omega X$ is positive definite and thus invertible. Deduce from this fact that $X^\top X$ is invertible.
2. Show that B is not necessarily invertible if we only assume that Ω is real, symmetric and invertible.

Solution

(a) Recall that X is full rank if and only if X is injective and if and only if $\ker(X) = \{0_p\}$. If $v \in \mathbb{R}^p \setminus \{0_p\}$,

$$v^\top B v = v^\top X^\top \Omega X v = (Xv)^\top \Omega X v > 0$$

since $Xv \neq 0_n$ and Ω is positive definite. It follows that B is also positive definite and thus invertible. The second part follows from the first upon taking $\Omega = I_n$, which is positive definite.

(b) Counter-example. With $X = (1,1)^\top$ and $\Omega = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$, we get $X^\top \Omega X = 0$. In general, if Ω has one positive eigenvalue a and one negative eigenvalue b, one can find a matrix X such that $X^\top \Omega X = 0$.

Problem 6. Let $Y_1, \ldots, Y_n$ be i.i.d. from $\mathcal{N}(\mu, \sigma^2)$.

1. Show that the log-likelihood satisfies

$$\ell(\mu, \sigma^2) = -\frac{1}{2}\left\{ n \log \sigma^2 + \frac{1}{\sigma^2} \sum_{j=1}^{n} (y_j - \mu)^2 \right\} + \text{const}$$

and the maximum likelihood (ML) estimates of μ and σ^2 are

$$\widehat{\mu} = \bar{y} \qquad \text{and} \qquad \widehat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} (y_j - \bar{y})^2.$$

2. Show that for μ fixed, the ML estimate of σ^2 is given by

$$\widehat{\sigma}^2_\mu = \frac{n-1}{n} s^2 \left\{ 1 + \frac{t(\mu)^2}{n-1} \right\},$$

where

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2 \qquad \text{and} \qquad t(\mu) = \sqrt{n}\, \frac{\bar{y} - \mu}{s}.$$

3. For a fixed μ, we take the ML estimate of σ^2 (depending on μ) and plug it in the likelihood to obtain the so-called profile likelihood for μ, which only depends on μ, not on σ. Show the profile likelihood is given by:

$$\ell_p(\mu) = -\frac{n}{2} \log[s^2\{1 + t(\mu)^2/(n-1)\}] + \text{const}.$$

Solution

1. An easy calculation.
2. To simplify the calculus, we denote $w = \sigma^2$. Then

$$\partial l / \partial w = -n/(2w) + \sum_{i=1}^{n} (y_i - \mu)^2/(2w^2).$$

The partial derivative is equal to zero at

$$w = n^{-1} \sum_i (y_i - \mu)^2. \tag{2}$$

To show the special form of $\widehat{\sigma}^2_\mu$, we substitute back for w:

$$\begin{aligned} n\widehat{\sigma}^2_\mu &= \sum_i (y_i - \bar{y} + \bar{y} - \mu)^2 \\ &= \sum_i (y_i - \bar{y})^2 + 2(\bar{y} - \mu) \underbrace{\sum_i (y_i - \bar{y})}_{=0} + \sum_i (\bar{y} - \mu)^2 \\ &= (n-1)s^2 + n(\bar{y} - \mu)^2 \\ &= (n-1)s^2 \left\{ 1 + \frac{1}{n-1} \left(\sqrt{n}\, \frac{\bar{y} - \mu}{s} \right)^2 \right\}. \end{aligned}$$

Upon noticing that the expression in parentheses is equal to $t(\mu)$, we have

$$\widehat{\sigma}^2_\mu = \frac{n-1}{n} s^2 \left\{ 1 + \frac{t(\mu)^2}{n-1} \right\}. \tag{3}$$

3. We use the symbol "$\propto$" to denote equality up to a constant. We have

$$\ell_p(\mu) \propto -(1/2) \left\{ n \log(\widehat{\sigma}^2_\mu) + \frac{1}{\widehat{\sigma}^2_\mu} \sum_i (y_i - \mu)^2 \right\}. \tag{4}$$

Plugging-in (2), we obtain

$$\frac{1}{\widehat{\sigma}^2_\mu} \sum_i (y_i - \mu)^2 = n.$$

Secondly, (3) gives us

$$\log(\widehat{\sigma}^2_\mu) = \log[(n-1)/n] + \log\left[s^2 \left\{ 1 + t(\mu)^2/(n-1) \right\} \right].$$

Plugging the previous two expressions back into (4), we have

$$\ell_p(\mu) \equiv -\frac{n}{2} \log\left[s^2 \left\{ 1 + t(\mu)^2/(n-1) \right\} \right].$$

Problem 7. Let Σ be an $p \times p$ positive definite covariance matrix. We define the precision matrix $Q = \Sigma^{-1}$. Suppose the matrices are partitioned into blocks,

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \text{ and } \Sigma^{-1} = Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}$$

with $\dim(\Sigma_{11}) = k \times k$ and $\dim(\Sigma_{22}) = (p-k) \times (p-k)$. Prove the following relationships

(a) $\Sigma_{12}\Sigma_{22}^{-1} = -Q_{11}^{-1}Q_{12}$

(b) $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = Q_{11}^{-1}$

(c) $\det(\Sigma) = \det(\Sigma_{22}) \det(\Sigma_{1|2})$ where $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Solution

By writing explicitly the relationship $Q\Sigma = I_n$, we get

$$\begin{aligned} Q_{11}\Sigma_{11} + Q_{12}\Sigma_{21} &= I_k \\ Q_{21}\Sigma_{12} + Q_{22}\Sigma_{22} &= I_{p-k} \\ Q_{21}\Sigma_{11} + Q_{22}\Sigma_{21} &= O_{p-k,k} \\ Q_{11}\Sigma_{12} + Q_{12}\Sigma_{22} &= O_{k,p-k}. \end{aligned}$$

Recall that we can only invert matrices whose double indices are identical and that both Q and Σ are symmetric, so $\Sigma_{12} = \Sigma_{21}^\top$. One easily obtains

(a) $\Sigma_{12}\Sigma_{22}^{-1} = -Q_{11}^{-1}Q_{12}$ making use of the last equation.

(b) $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = Q_{11}^{-1}$ by substituting Q_{12} from the last equation into the first.

(c) One can cleverly choose $B := \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ O & I \end{pmatrix}$, noting that $\det(B) = \det\left(B^\top\right) = 1$. Computing the quadratic form $B\Sigma B^\top$, we get $\det(\Sigma) = \det(\Sigma_{22})\det(\Sigma_{1|2})$ where $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Problem 8. Let $Y \sim \mathcal{N}_n(\mu, \Sigma)$ and consider the partition

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Y_1 is a $k \times 1$ and Y_2 is a $(n-k) \times 1$ vector for some $1 \le k < n$. Show that the conditional distribution of $Y_1 \mid Y_2 = y_2$ is $\mathcal{N}_k(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{1|2})$ and $\Sigma_{1|2}$ is the Schur complement of Σ_{22}.

Hint: write the joint density as $p(y_1, y_2) = p(y_1 \mid y_2)p(y_2)$ and express the joint density in terms of the precision matrix Q. It suffices to consider terms in $p(y_1, y_2)$ that depend only on y_1 (why?). The conditional distribution can then be identified by its functional form directly.

Solution

It is easier to obtain this result by expressing the density of the Gaussian distribution in terms of the precision matrix $Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}$ rather than in terms of the covariance matrix Σ.

Consider the partition $Y = (Y_1, Y_2)$. The conditional density as a function of y_1 is given, up to proportionality, by

$$\begin{aligned} f(y_1 \mid y_2) &\overset{y_1}{\propto} \exp\left(-\frac{1}{2}\left(y_1 - \mu_1\right)^\top Q_{11}\left(y_1 - \mu_1\right) - \left(y_1 - \mu_1\right)^\top Q_{12}\left(y_2 - \mu_2\right)\right) \\ &\overset{y_1}{\propto} \exp\left(-\frac{1}{2} y_1^\top Q_{11} y_1 - y_1^\top\left(Q_{11}\mu_1 - Q_{12}\left(y_2 - \mu_2\right)\right)\right) \end{aligned}$$

upon completing the square in y_1. This integrand is proportional to the density of a Gaussian distribution (and hence must be Gaussian) with precision matrix Q_{11}, while the mean vector and covariance matrix are

$$\begin{aligned} \mu_{1|2} &= \mu_1 - Q_{11}^{-1}Q_{12}\left(y_2 - \mu_2\right) \\ &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(y_2 - \mu_2\right) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

Note that $\Sigma_{1|2} = Q_{11}^{-1}$ corresponds to the Schur complement of Σ_{22}.

Remark: the above is sufficient (why?). The quadratic form appearing in the exponential term of the density of a Gaussian vector with mean ν and precision Ψ is

$$(x - \nu)^\top \Psi (x - \nu) = x^\top \Psi x - x^\top \Psi \nu - \nu^\top \Psi x + \nu^\top \Psi \nu.$$

uniquely determines the parameters of the Gaussian distribution. The quadratic term in x forms a sandwich around the precision matrix, while the linear term identifies the location vector. Since any (conditional) density function integrates to one, there is a unique normalizing constant and the latter need not be computed.

Alternative derivations of the result are given below. Suppose without loss of generality that $\mu = 0_n$. One can write the joint density as the product of the marginal and conditional densities

$$f_Y(y_1, y_2) = f_{Y_1|Y_2}(y_1 \mid y_2) f_{Y_2}(y_2).$$

Completing the square in y_1 in the exponential gives

$$y^\top Q y = (y_1 + Q_{11}^{-1}Q_{12}y_2)^\top Q_{11}(y_1 + Q_{11}^{-1}Q_{12}y_2) + y_2^\top \Sigma_{22}^{-1} y_2 \tag{S1}$$

and we can interpret the first term on the right hand side as a function of y_1 given y_2 and the second as y_2 alone. By using the transformation matrix $A = (I_k, O_{k,n-k})^\top$, we conclude that $y_2 \sim \mathcal{N}_{n-k}(\mu_2, \Sigma_{22})$. The marginal distribution can also be obtained by integrating out y_1 in $f_Y(y_1, y_2)$. Thus, the conditional density is the ratio

$$\begin{aligned}\frac{f_Y(y_1,y_2)}{f_{Y_2}(y_2)} &= \frac{(2\pi)^{-n/2}|Q|^{1/2}\exp\left(-\frac{1}{2}\left(y_1^\top Q_{11}y_1 + y_1^\top Q_{12}y_2 + y_2^\top Q_{21}y_1 + y_2^\top \Psi_{22}y_2\right)\right)}{(2\pi)^{-(n-k)/2}|\Sigma_{22}|^{-1/2}\exp\left(-\frac{1}{2}y_2\Sigma_{22}^{-1}y_2\right)} \\ &= (2\pi)^{-k/2}|Q|^{1/2}\frac{\exp\left(-\frac{1}{2}\left(y_1^\top Q_{11}y_1 + 2y_1^\top Q_{12}y_2 + y_2^\top Q_{21}Q_{11}^{-1}Q_{12}y_2\right)\right)}{\exp\left(-\frac{1}{2}\left(y_2\Sigma_{22}^{-1}y_2 + y_2^\top Q_{22}y_2 - y_2^\top Q_{21}Q_{11}^{-1}Q_{12}y_2\right)\right)}\end{aligned}$$

making use of the identity derived in 5.2 (c) to write $|Q| = |Q_{11}||\Sigma_{22}|^{-1}$. Using 5.2 (b) the denominator in the last line above is equal to one. The quadratic form in the exponential corresponds to the first term on the right hand side of (S1).

Alternatively, one could use the transformation matrix B from 7 (c) and look at $Z := By = (x^\top, Y_2^\top)^\top$ with Y assumed for now to have mean-zero. Since the Gaussian distribution is a location-scale family, the distribution of this affine transformation is $Z \sim \mathcal{N}_n(0, B\Sigma B^\top)$. We then have by 7 (c) that $Z \sim \mathcal{N}_n(0_n, \Omega)$ where

$$\Omega = \begin{pmatrix} Q_{11}^{-1} & O_{k,n-k} \\ O_{n-k,k} & \Sigma_{22} \end{pmatrix}.$$

It follows that X and Y_2 are independent, with $Y_2 \sim \mathcal{N}_{n-k}(0_{n-k}, \Sigma_{22})$ as given above. Now, $Y_1 = X + \Sigma_{12}\Sigma_{22}^{-1}Y_2$ thus conditional on $Y_2 = y_2$, we have $Y_1 \mid Y_2 \overset{\mathrm{d}}{=} X + \Sigma_{12}\Sigma_{22}^{-1}y_2$. It now suffices to note that for uncentered Y, we can write $(Y-\mu)+\mu$ with $Y-\mu$ centered. The expression given in the statement then follows immediately from the location change.

Problem 9. Let $Z \sim \mathcal{N}_n(0_n, I_n)$ and $Y \sim \mathcal{N}_n(\mu, \Sigma)$ with Σ positive definite.

(a) Let A be an orthogonal matrix. Show that $A^\top Z \sim \mathcal{N}_n(0_n, I_n)$.

(b) Show that $C^{-1}(Y-\mu) \sim \mathcal{N}_n(0_n, I_n)$ where C is the Cholesky root of Σ, the unique lower triangular matrix with positive diagonal elements such that $\Sigma = CC^\top$.

(c) Let H be a $n \times n$ projection matrix of rank $k \leq n$ with real entries. Show that $Z^\top HZ \sim \chi^2(k)$.

(d) Show that $(Y-\mu)^\top\Sigma^{-1}(Y-\mu) \sim \chi^2(n)$.

Solution

Recall the affine transformation property of the normal distribution:

$$Y \sim \mathcal{N}(\mu, \Sigma) \implies BY + \theta \sim \mathcal{N}(\theta + B\mu, B\Sigma B^\top). \tag{S1}$$

The Gaussian distribution is a location-scale family.

(a) Follows from (S1) and the fact that A is orthogonal, so $A^\top A = I_n$.

(b) The matrix C is invertible because its diagonal elements are all strictly positive. Since $C^{-1}(Y-\mu) = C^{-1}Y - C^{-1}\mu$, it follows from (S1) that $C^{-1}(Y-\mu)$ is normal with mean $C^{-1}\mu - C^{-1}\mu = 0_n$ and covariance $C^{-1}\Sigma C^{-\top} = C^{-1}CC^\top C^{-\top} = I_n$.

(c) Consider the spectral decomposition of the projection matrix

$$H = U\Lambda U^\top = \sum_{i=1}^{n}\lambda_i u_i u_i^\top = \sum_{i=1}^{k} u_i u_i^\top,$$

where without loss of generality the ith diagonal element of Λ, λ_i, is 1 if $i = 1, \ldots, k$ and zero otherwise. The first k columns of U form a basis for $\mathbb{R}^k$ and we can express HZ in this basis as $HZ = UX$, say. Since U is

orthogonal, one can write $X = \Lambda U^\top Z$ and so $X \sim \mathcal{N}_k(0_k, I_k)$. The result now follows from the definition of the $\chi^2(k)$ distribution.

(d) Since Σ is invertible it is positive definite. Write its Cholesky decomposition $\Sigma = CC^\top$, where C is invertible. From b), $Z := C^{-1}(Y-\mu) \sim \mathcal{N}_n(0_n, I_n)$ and

$$(Y-\mu)^\top \Sigma^{-1}(Y-\mu) = (Y-\mu)^\top C^{-\top} C^{-1}(Y-\mu) = Z^\top Z = Z^\top I_n Z.$$

The result now follows from c) since the identity matrix I_n is a projection matrix of rank n.

Problem 10. Consider a singular value decomposition (SVD) of the design matrix $X = UDV^\top$, where U is an $n \times p$ orthonormal matrix (meaning $U^\top U = I_p$ and the columns of U are orthogonal vectors), D is an $p \times p$ diagonal matrix and V is an $p \times p$ orthogonal matrix.

1. Show that H does not depend on V.
2. Give a formula for the ordinary least square estimate $\widehat{\beta}$, showing that the only inverse it involves is the inverse of a diagonal matrix.

Solution
The fact that both U and V are orthonormal means that $U^\top U = V^\top V = I_p$. The hat matrix is

$$H = UDV^\top (VDU^\top UDV^\top)^{-1} VDU^\top = U\Omega V^\top VD^{-2}V^\top VDU^\top = UU^\top$$

since $D = D^\top$ provided the diagonal matrix is square and $(VD^2V^\top)^{-1} = V^{-\top}D^{-2}V^{-1}$ where $V^{-1} = V^\top$.

If X is full rank, we can write

$$\widehat{\beta} = VD^{-1}U^\top y.$$

Problem 11. (Non-linear $\leftrightarrow$ linear models). This exercise has the goal of showing that a non-linear model can (sometimes) be transformed into a linear one. For instance, the model $y = \beta_1(x+\beta_3)^{\beta_2}(\varepsilon^2+1)$ can be written as

$$\log(y) = \underbrace{\log(\beta_1)}_{\beta_1^*} + \underbrace{\beta_2}_{\beta_2^*}\log(x+\beta_3) + \underbrace{\log(\varepsilon^2+1)}_{\varepsilon^*},$$

with β_3 fixed, and $\begin{bmatrix}1 & \log(x+\beta_3)\end{bmatrix}$ as design matrix. Moreover,we need $\beta_1 > 0, x+\beta_3 > 0$ in order to do the transformation.

Write, when possible, the following models as linear regressions, either by transforming and/or by fixing some parameters. Specify the new parameter (β^*), the new error (ε^*), restrictions (e.g. $\beta_1 > 0$) and give the design matrix, as in the example above:

a) $y = \beta_0 + \beta_1/x + \beta_2/x^2 + \varepsilon$

b) $y = \beta_0/(1+\beta_1 x) + \varepsilon$

c) $y = \beta_0/(\beta_1 x) + \varepsilon$

d) $y = 1/(\beta_0 + \beta_1 x + \varepsilon)$

e) $y = \beta_0 + \beta_1 x^{\beta_2} + \varepsilon$

f) $y = \beta_0 + \beta_1 x_1^{\beta_2} + \beta_3 x_2^{\beta_4} + \varepsilon$

g) $y = \beta_1 x_1^{\beta_2}\cos(x_2)^{\beta_3}\varepsilon$

h) $y = \beta_1 + x_1^{\beta_2}(2+\cos(x_2))^{\beta_3}(\varepsilon^2+1)$

Solution
Here is an example of solution (there could be others). The fixed parameters are underlined (e.g. $\underline{\beta_0}$).

a) $y = \begin{pmatrix}1 & \frac{1}{x} & \frac{1}{x^2}\end{pmatrix}(\beta_0 \quad \beta_1 \quad \beta_2)^\top + \varepsilon$

b) $y = (\frac{1}{1+\underline{\beta_1}x})(\beta_0) + \varepsilon$

Compiled 09/12/2019 at 18:51

c) $y = (1/x)(\gamma) + \varepsilon$ with $\gamma = \beta_0/\beta_1$ or $y = (\frac{1}{x\underline{\beta_1}})(\beta_0) + \varepsilon$

d) $1/y = (1 \quad x)(\beta_0 \quad \beta_1)^\top + \varepsilon$

e) $y = (1 \quad x^{\underline{\beta_2}})(\beta_0 \quad \beta_1)^\top + \varepsilon$

f) $y = (1 \quad x_1^{\beta_2} \quad x_2^{\beta_4})(\beta_0 \quad \beta_1 \quad \beta_3)^\top + \varepsilon$

g) $\log(y) = (1 \quad \log(x_1) \quad \log[\cos(x_2)])(\beta_1 \quad \beta_2 \quad \beta_3)^\top + \log(\varepsilon)$ with $x_1, \varepsilon > 0$ and $\cos(x_2) > 0$

h) $\log(y - \underline{\beta_1}) = (\log(x_1) \quad \log[2 + \cos(x_2)])(\beta_2 \quad \beta_3)^\top + \log(\varepsilon^2 + 1)$ with $x_1 > 0$.

Problem 12. Let $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1 \ldots, n$.

a) Write down the design matrix X. Calculate the elements of $X^\top X$, $X^\top Y$ and $(X^\top X)^{-1}$.

b) Show that $\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. How do you interpret the estimate?

Solution

a) One can straightforwardly calculate

$$X^\top X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}, \quad X^\top Y = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix}$$

and use the 2×2 matrix inversion formula to get

$$(X^\top X)^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

b) Formula for $\widehat{\beta}$ follows easily by multiplying $\widehat{\beta} = (X^\top X)^{-1} X^\top Y$, though we are only interested in the second element of the resulting vector.

Assume now that the data are standardized (i.e. both x and Y have mean zero and variance 1). Then $\widehat{\beta}_1$ reduces to the empirical correlation coefficient between x and Y, and it is the slope of the regression line when data are plotted. When we alleviate the assumption that our data are standardized, the interpretation of $\widehat{\beta}_1$ as the slope of the regression line is retained.

Problem 13. (Models in R)

In R, a model formula has the following general form

```
reponse~expression
```

where the left-hand side of the formula, `reponse`, can sometimes be absent, and the right-hand side, `expression`, is a collection of terms linked by operators, usually as an arithmetical expression. Let us suppose, for example, that

$$y = \begin{pmatrix} 217 \\ 143 \\ 186 \\ 121 \\ 157 \\ 143 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 1 & 3 \\ 0 & 2 & 1 \\ 1 & 2 & 2 \\ 0 & 2 & 3 \end{pmatrix},$$

and let `x`, `a`, `b` be the columns of $X = [x, a, b]$.

Compiled 09/12/2019 at 18:51

a) A *factor* is a variable that represents a categoric/qualitative variable (command `as.factor()` in R). For example, if `a` is a factor, then `y~a` represents the model

$$y_j = \beta_0 + \alpha_1 + \varepsilon_j, \quad j = 1,2,3; \qquad y_j = \beta_0 + \alpha_2 + \varepsilon_j, \quad j = 4,5,6.$$

Formally, it is written with indicators:

$$y_j = \beta_0 + \alpha_1 I_{(a_j = \text{“1”})} + \alpha_2 I_{(a_j = \text{“2”})} + \varepsilon_j, \tag{5}$$

where $I_E = 1$ if the expression E is true, and 0 otherwise. NoOtice that the “1” and “2” of “vector” `a` do *not* represent the number 1 and 2, but some categories, groups, classes or levels. For example, we can have “1” = “normally fed”, and “2” = “fed with growth inibitor”.

Let us suppose that `a` and `b` are factors:

I. Give the design matrix corresponding to model (5), as well as the vector of variables.

II. Notice that this matrix is *not* injective. What is the consequence on the parameters estimation?

III. Suppress the column corresponding to α_1 of this matrix in order to have an injective matrix. What is now the interpretation of the parameters β_0 and α_2?

IV. When the model includes the constant β_0, R automatically suppresses the first level of each factor. Give the design matrix corresponding to the following models:

(i) `y~a`, (ii) `y~a+b`, (iii) `y~x+a-1` (iv) `y~b+x-1`.

b) Let us suppose that `a` and `b` are (still) factors: a*interaction* is represented in the form `a:x` or `a:b`. For example, `y~a:x` represents

$$y_j = \beta_0 + \alpha_1 x_j + \varepsilon_j, \quad j = 1,2,3; \qquad y_j = \beta_0 + \alpha_2 x_j + \varepsilon_j, \quad j = 4,5,6;$$

that also writes

$$y_j = \beta_0 + \alpha_1 I_{(a_j = \text{“1”})} x_j + \alpha_2 I_{(a_j = \text{“2”})} x_j + \varepsilon_j$$

with the indicators. I.e., a model with different slopes for groups “1” and “2”, but with the same intercept.

Expression `y~a:b` represents the model

$$y_j = \beta_0 + \alpha_j + \varepsilon_j, \quad j = 1,\ldots,6;$$

that also writes

$$y_j = \beta_0 + \sum_{i=1}^{2}\sum_{l=1}^{3} \gamma_{i,l} I_{(a_j = \text{“}i\text{”})} I_{(b_j = \text{“}l\text{”})} + \varepsilon_j.$$

I.e., a model with different *intercepts* for different combinations of levels for `a` and `b`. Find the design matrices corresponding to models

(i) `y~a:x`, (ii) `y~a:b` (iii) `y~a+b:x`, (iv) `y~a+a:b:x`.

Among these matrices, which ones have linearly independent columns?

Solution

a) I. $X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$, $\beta = (\beta_0, \alpha_1, \alpha_2)^\top$.

II. $X(1,-1,-1)^\top = 0$, so it is not injecitve. The consequence is that we cannot invert $X^\top X$; statistically, it means that we cannot esetimate the three parameters at the same time. This model is not *identifiable*: `http://en.wikipedia.org/wiki/Identifiability`.

III. $X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$.

This corresponds to set $\alpha_1 = 0$. β_0 is the mean of each oservation in the group a_j = "1" and α_2 is the difference between the average of group a_j = "2" and the average of group a_j = "1".

IV.

$$X_{y\sim a} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, X_{y\sim a+b} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}, X_{y\sim x+a-1} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, X_{y\sim b+x-1} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

b)

$X_{y\sim a:x} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$; the columns are linearly independent.

$X_{y\sim a:b} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$; the columns are not linearly independent.

$X_{y\sim a+b:x} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$; the columns are linearly independent.

$X_{y\sim a+a:b:x} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$; the columns are not linearly independent.

c) You can verify your answers using the following commands:

```
> y <- c(217,143,186,121,157,143)
> X <- matrix(c(1,0,2,0,1,0,1,1,1,2,2,2,1,2,3,1,2,3), 6, 3)
```

```
> dfa <- data.frame(y = y, x = X[,1], a = X[,2], b = X[,3])
> dfb <- data.frame(y = y, x = X[,1], a = as.factor(X[,2]), b = as.factor(X[,3]))
> model.matrix(reponse~expression, data = dfa)
```

Problem 14. (Confounders and Simpson's paradox) In this exercise we are interested in the dependence of a standardized test *percentile* on the grade point average (*GPA*) of students of a certain high school in the US. The data file `percentile.RData` also contains the variable *grade*, which determines the study age of the students.

a) Load the data and create a scatterplot of *percentile* on *GPA*.

b) Fit the linear model `percentile~GPA` and add the regression line to your scatterplot from part a). What would be your conclusion about the relationship of *percentile* on *GPA* based on this model? How does the model quantify this relationship? Does this makes sense?

c) Add the variable *grade* to the model as a factor. How this changes your qualitative conclusions? How does the new model quantify the dependency? Are the conclusions sensible now?

d) Add the interaction term between *GPA* and *grade* to your model. What is changed compared to part c)?

Solution

The plots for every subquestion are given in the figure below.

a) The data are stored as an `.RData` file, hence it can be simply loaded as `load("percentile.RData")`. Then one can form the scatterplot using `plot(DATA\$percentile $\sim$ DATA\$GPA`.

b)
```
m1 <- lm(percentile ~ GPA, data=DATA)
summary(m1)
abline(m1$coefficients[1],m1$coefficients[2])
```

The previous commands gives tells us, that the correlation between *percentile* and *GPA* is negative (regression coefficient -3.773), but the relationship is not significant (t-test p-value 0.316). This means, that qualitatively there is no relationship between *percentile* and *GPA*. Quantitatively, one can say that improving a students *GPA* by 1 leads to decrease of his percentile by 3.773. This somehow counterintuitive. One would expect that better GPA should be associated with better percentile, provided that the education system is working.

c)
```
m2 <- lm(percentile ~ GPA+as.factor(grade), data=DATA)
summary(m2)
plot(DATA$percentile[DATA$grade==8] ~ DATA$GPA[DATA$grade==8],
     col="blue",xlim=c(1,4), ylim=c(10,100), main="c)")
points(DATA$percentile[DATA$grade==12] ~ DATA$GPA[DATA$grade==12],
       col="red",pch=0)
abline(m2$coefficients[1],m2$coefficients[2],col="blue")
abline(m2$coefficients[1]+m2$coefficients[3],m2$coefficients[2],col="red")
```

Once the variable *grade* is accounted for by the model, not only *GPA* becomes significant but the negative dependence from part b) suddenly becomes positive as one would expect. Since *grade* has only 2 levels (students are both from the 8th grade or 12th) it makes sense to treat different classes like two different groups (hence the coloring in the plot). Quantitatively, the model says that if student A has *GPA* larger by 1 than *GPA* of student B, student A's percentile is expected to be higher by 16.884 than that of student B.

d) The code here is only a slight modification of the previous one. Note that if one naturally wants both the interaction and the main terms, $*$ operator can be used as

```
m3 <- lm(percentile ~ GPA*as.factor(grade), data=DATA)
```

While the model from part c) only allowed for the intercept to be different for the two groups of students and the slope was fixed to be the same, model `m3` allows for both the intercept and the slope to be for the

two groups of students. Qualitative conclusions remain roughly the same, but one can notice that *GPA* has a slightly stronger effect among the 8th grade students. To put it in numbers, if student A has *GPA* larger by 1 than *GPA* of student B, student A's percentile is expected to be higher by 20.626 (respectively by 20.626-7.862=12.764) than that of student B in the case of both students being in the 8th grade (respectively the 12th grade).

Variable *grade* is the so-called confounder of the relationship between *percentile* and *GPA*. If *grade* is not accounted for, the model produces completely wrong results. In this case, including *grade* changes negative relationship to positive one, which is called the Simpson's paradox. Paradox, because even though that higher values of *GPA* are naturally associated with higher values of *percentile* in both of the two classes appearing in our data set, it seems at the first glance that the overall correlation between *GPA* and *percentile* is negative. A sensible explanation of this could be the following: younger students usually have better GPA's because they put more effort into the studies, but they are not yet educated enough to be able to score higher on a standardized test than their older colleagues.

Often, a confounder is not taken into account in a study, which then leads to insensible conclusions and subsequent tabloid headings such as "Want to go to Harvard? Fail high school first!".

Problem 15. Assume a linear model was developed for blood glucose concentration (Y) of a patient after giving u units of a medicament to the patient with weight w and sex g (0=male, 1=female). In this model, the effect of weight w and the medicament dose u on the glucose concentration Y is different for males and females. Contrarily, the increase of the medicament dose u by 1 has (for two patients of the same sex and weight) the same effect on Y regardless the (actual value of) weight of the patient.

a) Write down the regression function of the model.

b) Assume the first observation is based on a male, 80 kg, who was given 10 units of the medicament. The second observation is based on a female, 60 kg, who was give 8 units of the medicament. Write down the first two rows of the design matrix.

c) How would you test whether weight w has different effect on Y based on the sex g?

Solution

The solutions are not unique, they depend on the ordering of variables.

a) A possible regression function can be

$$\mathsf{E}(Y) = \beta_0 + \beta_1 u + \beta_2 w + \beta_3 g + \beta_4 ug + \beta_5 wg.$$

Note that since g attains only two values, it does not matter in this case whether it is considered as a factor or as a continuous variable. But it would be more natural to consider it as a factor:

$$\mathsf{E}(Y) = \beta_0 + \beta_1 u + \beta_2 w + \beta_3 \delta_{[g=1]} + \beta_4 u\delta_{[g=1]} + \beta_5 w\delta_{[g=1]},$$

where δ is the identifier operator.

b) The following matrix is the design matrix corresponding to the regression function from part a):

$$X = \begin{pmatrix} 1 & 10 & 80 & 0 & 0 & 0 \\ 1 & 8 & 60 & 1 & 8 & 60 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

c) One would test $H_0 : \beta_5 = 0$, preferably by the F-test.

Problem 16. Suppose the $n \times p$ full-rank design matrix X can be partitioned into two blocks as $[X_1\ X_2]$ and let $M_{X_1} := I_n - H_{X_1}$. Show that $H_X = H_{X_1} + H_{M_{X_1}X_2}$, where $H_{M_{X_1}X_2}$ is the projection on to the span of $M_{X_1}X_2$. (Draw a 3D picture to visualize what this result actually says.)

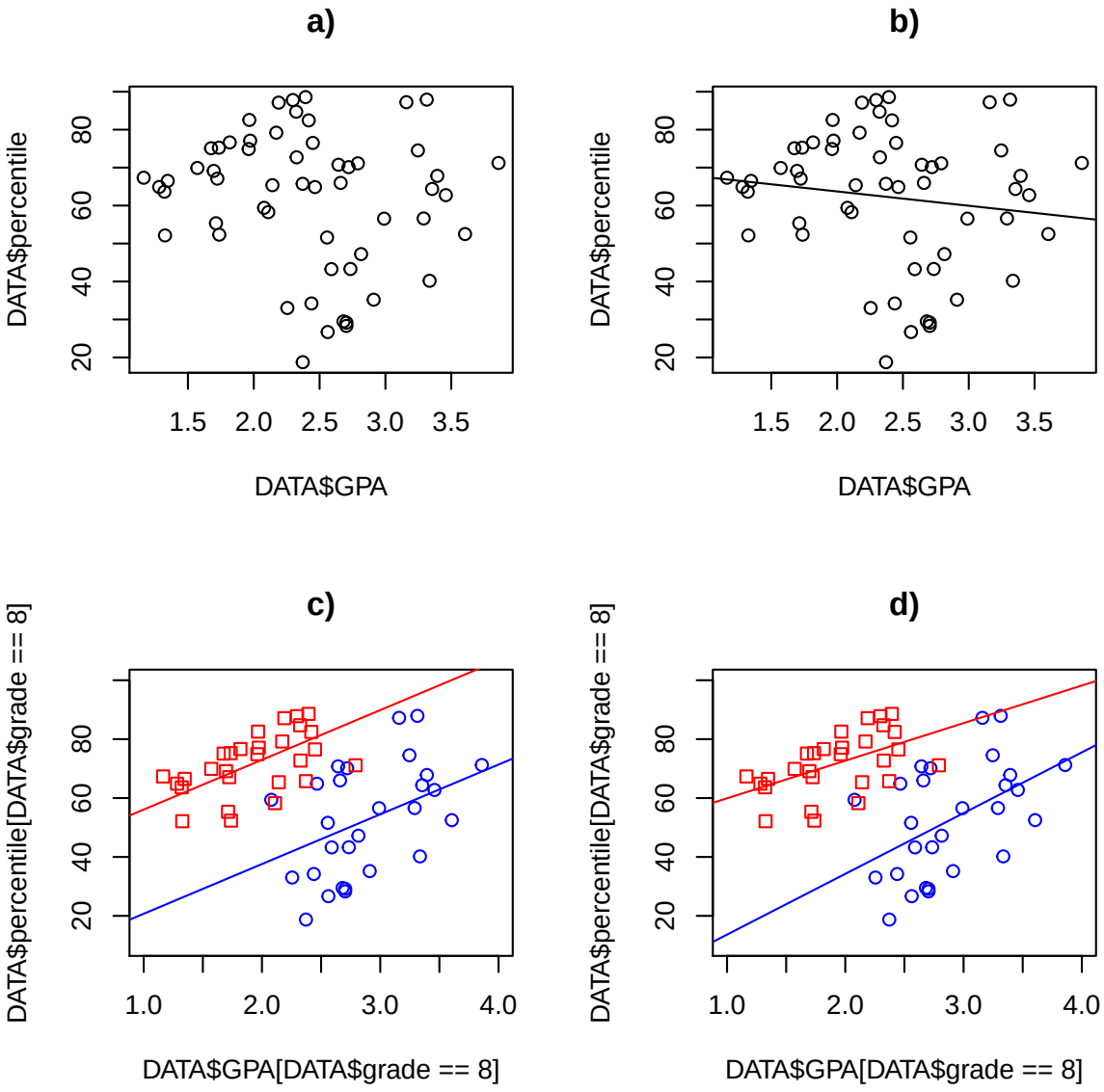

Figure 1: Standardised residuals as function of values adjusted for four Gaussian models.

Solution

We need to show that $H_{X_1} + H_{M_{X_1}X_2}$ is an orthogonal projection matrix, i.e., it is idempotent, symmetric and it spans $\mathcal{S}(X)$. Note that $X_1^\top M_{X_1} X_2 = O$, so $H_{X_1} H_{M_{X_1}X_2} = O$ also. Since both $H_{M_{X_1}X_2}$ and H_{X_1} are orthogonal projection matrices, the first two statements are obvious.

It remains to show that any vector $z \in \mathcal{S}(X)$ is invariant under the action of $H_{X_1} + H_{M_{X_1}X_2}$ and that any vector orthogonal to this span is annihilated by $H_{X_1} + H_{M_{X_1}X_2}$. Since X is full rank, we can write $z = X\gamma = X_1\gamma_1 + X_2\gamma_2$ for some vector γ and subvectors γ_1 and γ_2. Then

$$\begin{aligned}(H_{X_1} + H_{M_{X_1}X_2})z &= (H_{X_1} + H_{M_{X_1}X_2})(X_1\gamma_1 + X_2\gamma_2)\\ &= H_{X_1}(X_1\gamma_1 + X_2\gamma_2) + H_{M_{X_1}X_2}(X_1\gamma_1 + X_2\gamma_2)\\ &= X_1\gamma_1 + H_{X_1}X_2\gamma_2 + M_{X_1}X_2\gamma_2\\ &= X_1\gamma_1 + X_2\gamma_2\end{aligned}$$

upon noting that

$$\begin{aligned}H_{M_{X_1}X_2}X_1 &= M_{X_1}X_2(X_2^\top M_{X_1}X_2)^{-1}X_2^\top M_{X_1}X_1 = O,\\ H_{M_{X_1}X_2}X_2 &= M_{X_1}X_2(X_2^\top M_{X_1}X_2)^{-1}X_2^\top M_{X_1}X_2 = M_{X_1}X_2.\end{aligned}$$

Take now $w \in \mathcal{S}^\perp(X)$. We have

$$\begin{aligned}(H_{X_1} + H_{M_{X_1}X_2})w &= H_{X_1}w + H_{M_{X_1}X_2}w\\ &= 0 + M_{X_1}X_2(X_2^\top M_{X_1}X_2)^{-1}X_2^\top M_{X_1}w\\ &= M_{X_1}X_2(X_2^\top M_{X_1}X_2)^{-1}X_2^\top(I - H_{X_1})w = 0.\end{aligned}$$

Indeed, $H_{X_1}w = 0$ because w is orthogonal to X, thus also orthogonal to X_1. At the same time, $X_2^\top w = 0$ by orthogonality. By uniqueness of projection matrices, the result follows.

Problem 17. (Forecast and confidence intervals).

The following table gives the estimations, the standardised errors and the correlations for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ adjusted for $n = 13$ cement data of the example given at course.

```
             Estimate         SE           Correlations of Estimates
(Intercept)    48.19      3.913             (Intercept)    x1    x2
        x1      1.70      0.205         x1     -0.736
        x2      0.66      0.044         x2     -0.416   -0.203
        x3      0.25      0.185         x3     -0.828    0.822  -0.089
```

a) Explain how we can compute the standardised errors and correlations in the table above.

b) For this model, what is the forecast of y for $x_1 = x_2 = x_3 = 1$? How much would the prediction increase if $x_1 = 5$? And if $x_1 = x_2 = 5$?

c) For this model, compute, using only the information above and the fact that $t_9(0.975) = 2.262$ and $t_9(0.95) = 1.833$, the 0.95 confidence intervals for β_0, β_1, β_2 and β_3. Compute also a 0.90 confidence interval for $\beta_2 - \beta_3$.

Solution

a) The covariance of $\widehat{\beta}$ is given by $\mathsf{Var}\widehat{\beta} = \sigma^2(X^\top X)^{-1}$. Since we do not know σ^2, we estimate the covariance with $\widehat{\mathrm{var}}(\widehat{\beta}) = S^2(X^\top X)^{-1}$. Denoting $v_{ij} = ((X^\top X)^{-1})_{ij}$, $i = 0,1,2,3$, $j = 0,1,2,3$ (note that we start by the 0 index). Hence, the i-th standardised error is estimated by $\widehat{\mathrm{SE}}(\widehat{\beta}_i) = \sqrt{\widehat{\mathrm{var}}(\widehat{\beta})_{ii}} = \sqrt{S^2 v_{ii}}$. For the correlation, we have

$$\widehat{\mathrm{corr}}(\widehat{\beta}_i, \widehat{\beta}_j) = \frac{\widehat{\mathrm{var}}(\widehat{\beta})_{ij}}{\sqrt{\widehat{\mathrm{var}}(\widehat{\beta})_{ii}}\sqrt{\widehat{\mathrm{var}}(\widehat{\beta})_{jj}}} = \frac{S^2 v_{ij}}{\sqrt{S^2 v_{ii}}\sqrt{S^2 v_{jj}}} = \frac{v_{ij}}{\sqrt{v_{ii}v_{jj}}}.$$

b) We recall that the forecast is given by

$$\widehat{y}_+ = x_+^\top \widehat{\beta}.$$

Here, we have

$$\widehat{y}_+ = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3.$$

For $x_1 = x_2 = x_3 = 1$, the expectation would increase of $4\widehat{\beta}_1 = 4 \times 1.70 = 6.80$ if $x_1 = 5$, and of $4\widehat{\beta}_2 = 4 \times 0.66 = 2.64$ if $x_2 = 5$. Explicitly,

$$\begin{aligned}
x_1 = x_2 = x_3 = 1 &\implies \widehat{y}_+ = \widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3 = 48.19 + 1.70 + 0.66 + 0.25 = 50.80 \\
x_1 = 5, x_2 = x_3 = 1 &\implies \widehat{y}_+ = 48.19 + 1.70 \times 5 + 0.66 + 0.25 = 57.60 \\
x_1 = x_2 = 5, x_3 = 1 &\implies \widehat{y}_+ = 48.19 + 1.70 \times 5 + 0.66 \times 5 + 0.25 = 60.24
\end{aligned}$$

c) Let us denote here $(X^\top X)^{-1} = (v_{ij})_{i,j=0}^3$. The entries v_{ij} can be read out of the R output provided in the assignment.

Recall that for the i-th coordinate of β, the confidence interval is

$$\widehat{\beta_i} \pm \sqrt{S^2 v_{ii}}\, t_{n-p}(\alpha/2) = \widehat{\beta_i} \pm \widehat{\mathrm{SE}}(\widehat{\beta}_i)\, t_{n-p}(\alpha/2), \quad i = 0,1,2,3.$$

Here, $n = 13$, $p = 4$, $\alpha = 0.05$, $t_9(0.975) = 2.262$, so we have the four intervals:

$$[39.34, 57.04], \quad [1.236, 2.164], \quad [0.5605, 0.7595], \quad [-0.1685, 0.6685].$$

For the test $\beta_3 = 0$ against $\beta_3 \neq 0$ we do not reject the null hypothesis because $0 \in [-0.1685, 0.6685]$.

More generally, if $c \in \mathbb{R}^p$, the confidence interval for $c^\top \beta$ is given by

$$c^\top \widehat{\beta} \pm t_{n-p}(\alpha/2) \sqrt{S^2 c^\top (X^\top X)^{-1} c}.$$

Here we want a confidence interval for $c^\top \beta$ with $c = (0,0,1,-1)^\top$. We find

$$\begin{aligned}
S^2 c^\top (X^\top X)^{-1} c &= S^2 v_{22} + S^2 v_{33} - 2 \frac{v_{23}}{\sqrt{v_{22} v_{33}}} \sqrt{S^2 v_{22}} \sqrt{S^2 v_{33}} \\
&= \left(\widehat{\mathrm{SE}}(\widehat{\beta}_2)\right)^2 + \left(\widehat{\mathrm{SE}}(\widehat{\beta}_3)\right)^2 - 2\, \widehat{\mathrm{corr}}(\widehat{\beta}_2, \widehat{\beta}_3)\, \widehat{\mathrm{SE}}(\widehat{\beta}_2)\, \widehat{\mathrm{SE}}(\widehat{\beta}_3) \\
&= 0.044^2 + 0.185^2 - 2 \cdot (-0.089) \cdot 0.044 \cdot 0.185
\end{aligned}$$

Thus we have

$$[0.66 - 0.25 \pm \left\{0.044^2 + 0.185^2 - 2 \cdot 0.044 \cdot 0.185 \cdot (-0.089)\right\}^{1/2} t_9(0.95)] = [0.055, 0.765]$$

as 0.90 confindence interval for $\beta_2 - \beta_3$.

In R,

```
library(MASS)
fit<-lm(y~1+x1+x2+x3, data=cement)
confint(fit)
```

donne

```
2.5 %     97.5 %
(Intercept) 39.3411244 57.0461442
x1           1.2330935  2.1586869
x2           0.5568501  0.7569797
x3          -0.1678276  0.6678628
```

for the confidence intervals of each coordinate of β.

Problem 18. (Linear Gaussian models and space rotations) Let

$$Y = X\beta + \varepsilon,$$

be a Gaussian linear model, where X is injective, and $\varepsilon \sim N(0, \sigma^2 I)$. We know that if A is an orthogonal matrix, then $\widetilde{Y} = AY$ follows a linear Gaussian model as well,

$$\widetilde{Y} \sim \mathcal{N}(\widetilde{X}\beta, \sigma^2 I),$$

with $\widetilde{X} = AX$. We will consider some particular case the the orthogonal matrix A:

I. $A = U^\top$, where $X = U\Lambda V^\top$ is the singular values decomposition of X.

II. $A = Q^\top$, where $X = QR$ is the QR decomposition of X

For each of these cases,

a) Compute the adjusted values $\widehat{\widetilde{y}}$ as functions of $\widetilde{y}$. What can we say about their first p coordinates? And about their last $n-p$ coordinates?

b) Compute the residuals of model $\widetilde{Y}$. What can we say about their first p residuals? And about their last $n-p$ residuals?

c) Recall that residuals are usually dependent. What do we notice here?

Hint: Start by computing the *hat matrix* $\widetilde{H}$ for both cases I. and II.

Solution (a))
Let us compute $\widetilde{H}$ for each case:

I. The singular values decomposition of $X_{n\times p}$ is $U\Lambda V^\top$, with $\Lambda_{n\times p}$ diagonal, i.e.,

$$\Lambda = \begin{pmatrix} \Lambda_1 \\ 0 \end{pmatrix},$$

where Λ_1 is a $p \times p$ diagonal matrix. Since $\widetilde{X} = AX = \Lambda V^\top$,

$$\widetilde{X}^\top \widetilde{X} = V\Lambda_1^2 V^\top,$$

and its inverse is given by $V\Lambda_1^{-2}V^\top$ (Λ_1 is invertible since X is injective) and

$$\widetilde{H} = \widetilde{X}(\widetilde{X}^\top\widetilde{X})^{-1}\widetilde{X}^\top = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}.$$

Compiled 09/12/2019 at 18:51

II. Since $\widetilde{X} = R = (R_1, 0)^\top$, we have

$$\widetilde{H} = \widetilde{X}(\widetilde{X}^\top \widetilde{X})^{-1}\widetilde{X}^\top = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence, in the two cases

$$\widehat{\widetilde{y}} = (\widetilde{y}_1, \dots, \widetilde{y}_p, 0, \dots, 0)^\top$$

and

$$\widetilde{e} = (0, \dots, 0, \widetilde{y}_{p+1}, \dots, \widetilde{y}_n)^\top.$$

The first p coordinates of $\widehat{\widetilde{y}}$ are equal to those of $\widetilde{y}$, the last $n-p$ are zeros. The first p coordinates of $\widetilde{e}$ are zeros, and its last $n-p$ coordinates are $\widetilde{y}_i$, $i = n-p, \dots, n$. De plus,

$$\widetilde{e} = (I - \widetilde{H})\widetilde{y} \sim \mathcal{N}\left((I-\widetilde{H})\widetilde{X}\beta, (I-\widetilde{H})\sigma^2 I(I-\widetilde{H})\right) = \mathcal{N}\left(0, \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2 I_{n-p} \end{pmatrix}\right),$$

and thus the residuals are independent in this case (indeed, the first p are all 0 and the last $n-p$ are all i.i.d. Gaussians). Notice that, usually, the residuals are not independent!

Problem 19. (The best design)

Let us consider the simple regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\beta_0, \beta_1 \in \mathbb{R}$, $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{var}(\varepsilon) = \sigma^2 I_n$ (and $n \geq 2$).

a) Find the design matrix corresponding to this model and give a necessary and sufficient condition for it to be full rank.

b) Find the covariance matrix of the least squares estimator $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)^\top$.

c) Let us suppose that we can design the experiment by choosing $x_i \in [-1, 1]$ arbitrarily. Which is the best choice of x_i that minimises the variance of $\widehat{\beta}_1$?

Solution

a) The model can be written as

$$y = [X]\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \varepsilon, \quad \text{with } X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

The necessary and sufficient condition for the matrix X to be full rank is that the x_i's are not all the same.

b) From the model assumption we know that $\mathrm{var}(y) = \sigma^2 I_n$. We recall that $\widehat{\beta}$ is a linear transformation of y, i.e. $\widehat{\beta} = Ay$ with $A = (X^\top X)^{-1}X^\top$. Thus (recall that $X^\top X$ is symmetric, so $(X^\top X)^\top = X^\top X$)

$$\begin{aligned} \mathrm{var}(\widehat{\beta}) &= A\,\mathrm{var}(y)\,A^\top \\ &= \sigma^2 (X^\top X)^{-1} X^\top \left[(X^\top X)^{-1} X^\top\right]^\top \\ &= \sigma^2 (X^\top X)^{-1} X^\top X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1} \end{aligned}$$

c) The variance of $\widehat{\beta}_1$ is the second diagonal element of the variance matrix of $\widehat{\beta}$ which is

$$\mathrm{var}(\widehat{\beta}_1) = [\mathrm{var}(\widehat{\beta}_1)]_{22} = \sigma^2 [(X^\top X)^{-1}]_{22} = \frac{\sigma^2}{\det(X^\top X)} [X^\top X]_{11}.$$

From the form of X we have

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \implies X^\top X = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}.$$

So, the determinant of $X^\top X$ as a function of $x = (x_1, \ldots, x_n)^\top$ is

$$f(x) = n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2.$$

Summarizing, we have the dependence of $\mathrm{var}(\widehat{\beta}_1)$ as a function of x is

$$\mathrm{var}(\widehat{\beta}_1)(x) = \frac{\sigma^2 n}{f(x)},$$

and finding the $\arg\min \mathrm{var}(\widehat{\beta}_1)(x)$ is equivalent to finding the $\arg\max f(x)$. The only stationary point of $f(x)$ is a minimum (attained when $x_i = c$ for any $i = 1, \ldots, n$). Hence the maximum is attained on the boundary of the domain $[-1,1]^n$, i.e. for $x_i \in \{-1, 1\}$. As a consequence, $\sum_{i=1}^n x_i^2 = n$ and $\sum_{i=1}^n x_i = n_+ - n_-$, where n_+ is the number of x_i's attaining the value $+1$ and n_- is the number of x_i's attaining the value -1. When n is even the optimal value can be attained for $n_+ = n_- = n/2$, so $f(x) = n^2$ and $\mathrm{var}(\widehat{\beta}_1) = \sigma^2/n$. When n is odd we have a sub-optimal case and the maximum value is attained for $n_+ = n_- = (n-1)/2$ and exactly one $x_i = 0$, so $f(x) = n(n-1)$ and $\mathrm{var}(\widehat{\beta}_1) = \sigma^2/(n-1)$.

We can interpret the result in the following way: $\widehat{\beta}_1$ is the slope of the line that best fits the data according to the linear regression. If all values of x_i are close to a single value (say 0) there will be "many" accettably good linear fitting of the data and the slope can take values in a large set of values. Alternatively, small changes in the values of the y_i's can lead to large changes in the slope of the fitting line. On the contrary, if the x_i's are as spread as possible then even large changes in the values of the y_i's will have little effect on the value of the slope of the fitting line.

Problem 20. (Reformulation of the Gauss-Markov theorem)

Let $Y = X\beta + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0, \mathrm{var}(\varepsilon) = \sigma^2 I$. Let $\widehat{\beta}$ be the least squares estimator of β, and $\widetilde{\beta}$ another *linear* and *unbiased* estimator of β.

Show that

$$\mathrm{MSE}(c^\top \widetilde{\beta}) \geq \mathrm{MSE}(c^\top \widehat{\beta}), \quad \forall c \in \mathbb{R}^p,$$

is equivalent to the conclusion of Gauss-Markov theorem. Here, $\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}((\widehat{\theta} - \theta)^2)$ is the mean square error of $\widehat{\theta}$.

Recall: $\mathrm{MSE}(\widehat{\theta}) = \mathrm{bias}(\widehat{\theta})^2 + \mathrm{var}(\widehat{\theta})$.

Solution

We have

$$\begin{aligned} \mathrm{MSE}(c^\top \widetilde{\beta}) &= \underbrace{\mathrm{bias}(c^\top \widetilde{\beta})^2}_{=0, \text{ as } \widetilde{\beta} \text{ unbiased}} + \mathrm{var}(c^\top \widetilde{\beta}) = c^\top \mathrm{var}(\widetilde{\beta}) c, \\ \mathrm{MSE}(c^\top \widehat{\beta}) &= \underbrace{\mathrm{bias}(c^\top \widehat{\beta})^2}_{=0, \text{ as } \widehat{\beta} \text{ unbiased}} + \mathrm{var}(c^\top \widehat{\beta}) = c^\top \mathrm{var}(\widehat{\beta}) c. \end{aligned}$$

So

$$\mathrm{MSE}(c^\top \widetilde{\beta}) - \mathrm{MSE}(c^\top \widehat{\beta}) = c^\top \mathrm{var}(\widetilde{\beta}) c - c^\top \mathrm{var}(\widehat{\beta}) c = c^\top (\mathrm{var}(\widetilde{\beta}) - \mathrm{var}(\widehat{\beta})) c.$$

Hence, we have

$$\begin{aligned} & \mathrm{MSE}(c^\top \widetilde{\beta}) \geq \mathrm{MSE}(c^\top \widehat{\beta}), \quad \forall c \in \mathbb{R}^p \\ \Leftrightarrow \quad & \mathrm{MSE}(c^\top \widetilde{\beta}) - \mathrm{MSE}(c^\top \widehat{\beta}) \geq 0, \quad \forall c \in \mathbb{R}^p \\ \Leftrightarrow \quad & c^\top (\mathrm{var}(\widetilde{\beta}) - \mathrm{var}(\widehat{\beta}))c \geq 0, \quad \forall c \in \mathbb{R}^p \\ \Leftrightarrow \quad & \mathrm{var}(\widetilde{\beta}) - \mathrm{var}(\widehat{\beta}) \geq 0. \end{aligned}$$

Problem 21. (Diagnostic's graphics)

a) Figure 2 represents the standardised residuals as function of values adjusted for the linear model derived from four different dataset. For each case, discuss the adjusting and explain briefly how would you try to rememdy the possible insufficency.

b) Figure 3 shows four Q-Q Gaussian plots. In all the cases, the data do not follow the gaussian distribution. In fact, the data are generated from a distribution with

   i) tails heavier than Gaussian tails;

   ii) tails lighter than Gaussian tails;

   iii) a positive skewness coefficient;

   iv) a negative skewness coefficient.

   Associate each case i)–iv) with a Q-Q plot of Figure 3.

Solution

a) We know that $\mathrm{cov}(e, \widehat{y}) = 0$ and that the standardised residuals are standard gaussian random variables (i.e. residuals must take values between -2 and 2 independently from the values of $\widehat{y}_j$), if the starting hypotheses are respected ($\epsilon \sim N(0, \sigma^2 I), \ldots$).

   - Plot A : OK.
   - Plot B : Problem = An outlier.
   - Plot C : Problem = dependence between the fitted values and the standardised residuals. (see Example 8.24, page 390, Statistical Models, Davison).
   - Plot D : Problem = The variance of the residuals is not constant„ heteroscedasticity.

b)
   - Plot A: negative skewness coefficient.
   - Plot B: tails lighter than Gaussian tails.
   - Plot C: tails heavier than Gaussian tails.
   - Plot D: positive skewness coefficient.

Problem 22. (QQ plots)

The goal of this exercise is to justify the use of QQ plot to "see" whether a sample $x_1, \ldots, x_n$ comes from the normal distribution. Let $X_1, \ldots, X_n \sim N(0,1)$ be i.i.d, and let Φ be the cumulative distribution function of the normal law $N(0,1)$.

1. Show that $\Phi(X_1), \ldots, \Phi(X_n) \sim U([0,1])$ is i.i.d., where $U([0,1])$ denotes the uniform law on $[0,1]$.

2. Let $V_1, \ldots, V_n \sim U([0,1])$ be i.i.d., and let

$$V_{(1)} \leq V_{(2)} \leq \cdots \leq V_{(n)}$$

   be the associated order statistics. Compute the expectation of $V_{(k)}$.

   Hint: Use the fact f_k is a density function.

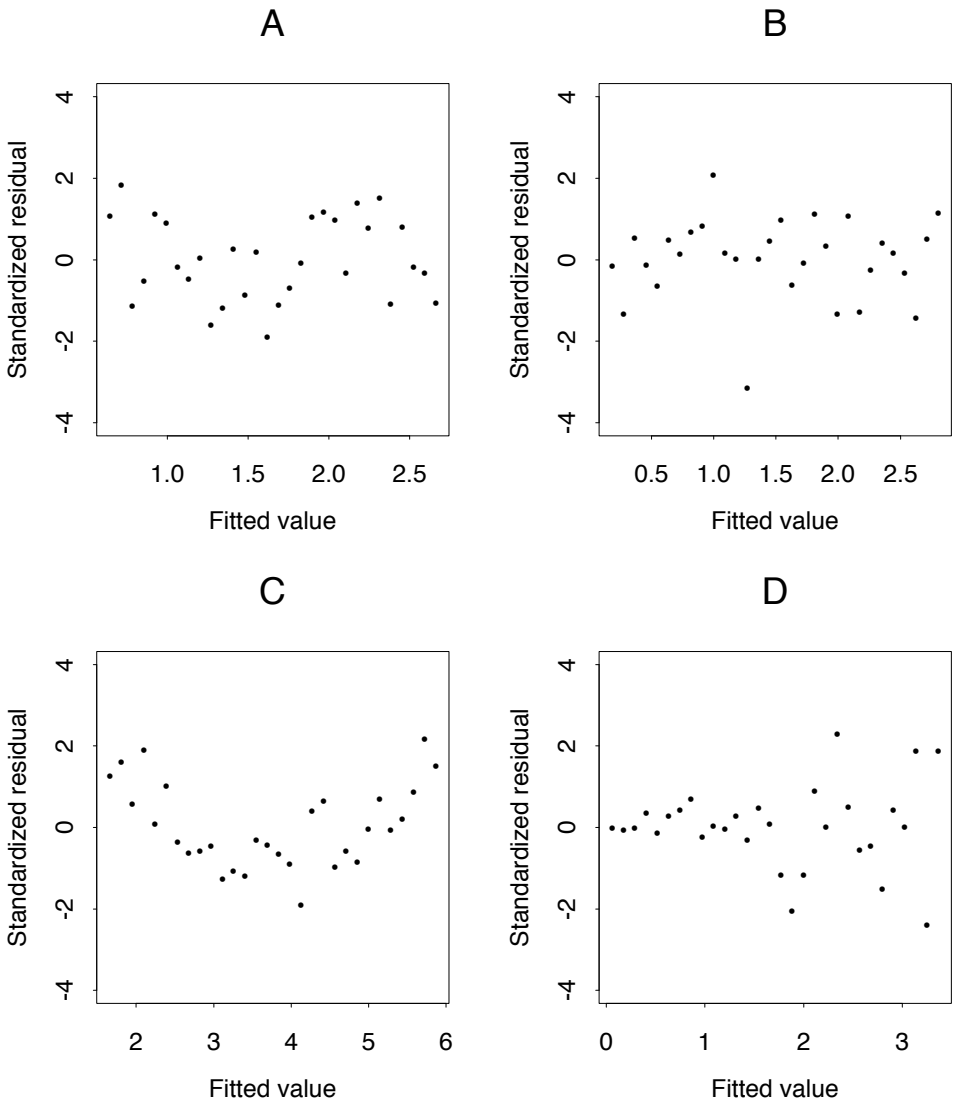

Figure 2: Standardised residuals as function of values adjusted for four Gaussian models.

Compiled 09/12/2019 at 18:51

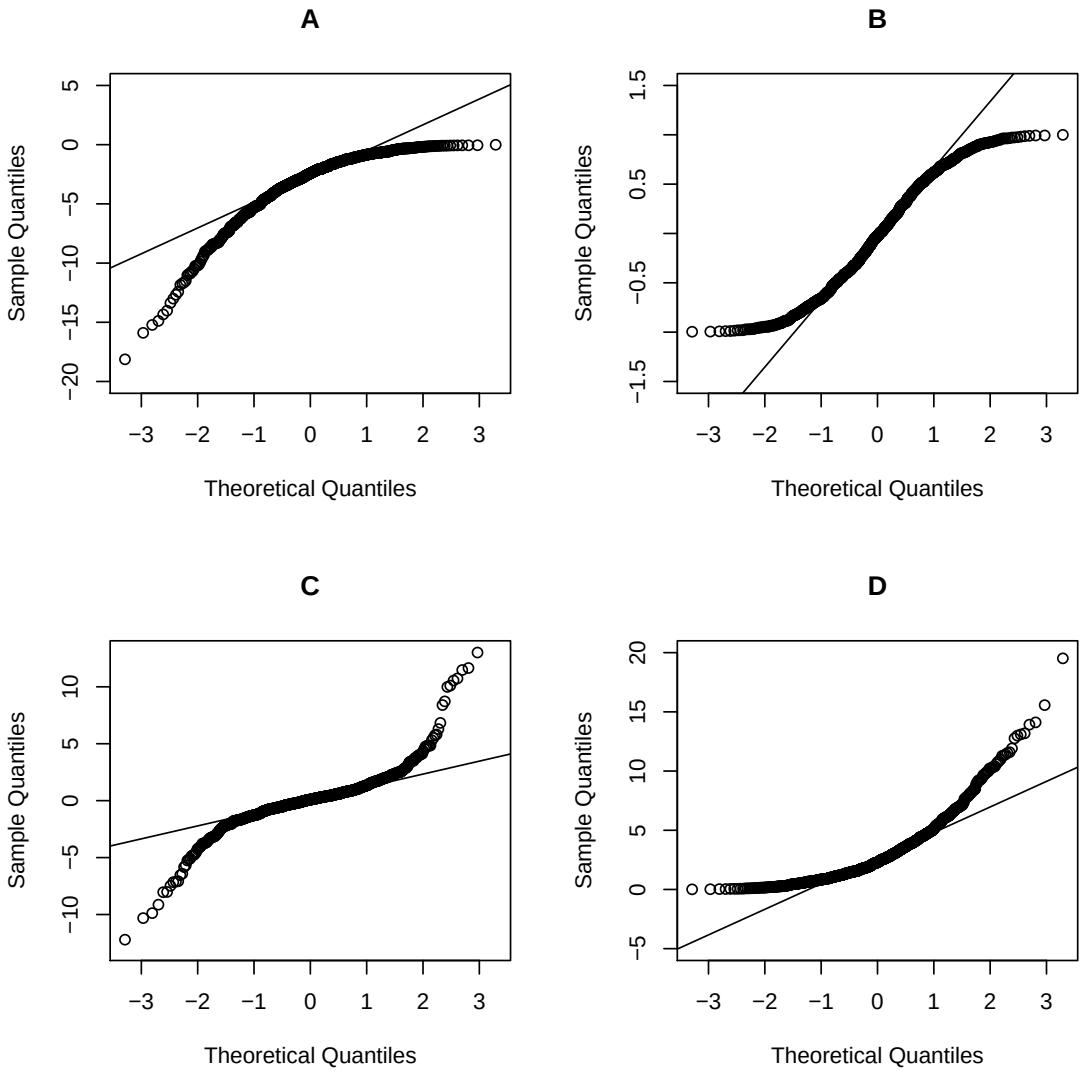

Figure 3: Four Q-Q Gaussian plots where the data do not follow a Gaussian law.

Compiled 09/12/2019 at 18:51

3. Let z_α be the quantile α of the normal law $N(0,1)$, defined by

$$\Phi(z_\alpha) = \alpha.$$

Explain why $\mathbb{E}[X_{(k)}] \approx z_{k/(n+1)}$. A rigorous justification is *not* necessary. Link it with the QQ plot.

Tip: It is necessary to approximate $\mathbb{E}[f(X)] \approx f(\mathbb{E}[X])$ for a function f slightly non linear.

4. ***Bonus:*** Prove that $V_{(k)} \sim \text{Beta}(k, n+1-k)$ with probability density function:

$$f_k(x) = n\binom{n-1}{k-1} x^{k-1}(1-x)^{n-k}, \quad x \in [0,1].$$

Attention: Even if there are not many calculations, it is not an easy exercise.

Tip: Let $A = \{0 < v_1 < \cdots < v_n < 1\} \subset [0,1]^n$. For $(v_1, \ldots, v_n) \in A$, use the symmetry of the problem to write

$$\mathbb{P}\left(V_{(1)} \le v_1, \ldots, V_{(n)} \le v_n\right)$$

as a n variables multiple integral. It is not advisable to compute explicitly this integral, but we can find a (very!) easy explicit formula for the joint distribution

$$\frac{\partial^n}{\partial v_1 \ldots \partial v_n} \mathbb{P}\left(V_{(1)} \le v_1, \ldots, V_{(n)} \le v_n\right).$$

Then, the marginal density of $V_{(k)}$ is found by integration the joint density over all other variables.

Solution

1. $x \mapsto \Phi(x)$ is strictly increasing, and $\Phi(\mathbb{R}) = (0,1)$. So, $\Phi(X_i) \in (0,1)$. If $x \in (0,1)$, then

$$P(\Phi(X_i) < x) = P(X_i < \Phi^{-1}(x)) = \Phi(\Phi^{-1}(x)) = x$$

hence $\Phi(X_i) \sim U([0,1])$.

2. Here $f_k(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}$, for $x \in [0,1]$, and zero otherwise. Since f_k is a density, $1 = \int f_k(x)dx$, (here and then, $\int = \int_0^1$) and thus

$$\int x^{k-1}(1-x)^{n-k} dx = \frac{(k-1)!(n-k)!}{n!},$$

or, more explicitly:

$$\int x^a (1-x)^b dx = \frac{a!b!}{(a+b+1)!}, \quad a, b \in \mathbb{N}.$$

Hence

$$\mathbb{E}[V_{(k)}] = \int x f_k(x) dx = \frac{n!}{(k-1)!(n-k)!} \int x^k (1-x)^{n-k} dx = \cdots = k/(n+1).$$

3. $\mathbb{E}[X_{(k)}] = \mathbb{E}[\Phi^{-1}(\Phi(X_{(k)}))] \approx \Phi^{-1}(\mathbb{E}[\Phi(X_{(k)})]) = \Phi^{-1}(k/(n+1)) = z_{k/(n+1)}$. Thus, when $X_1, \ldots, X_n$ are $N(0,1)$ i.i.d, we expect that their QQ normal plot is, "on average", on the line $y = x$.

4. We start by computing $\mathbb{P}\left(V_{(1)} \le v_1, \ldots, V_{(n)} \le v_n\right)$ under the assumption that $(v_1, \ldots, v_n) \in A$. The event $\{V_{(1)} \le v_1, \ldots, V_{(n)} \le v_n\}$ equals the union of all possible permutations of the event $\{V_1 \le v_1, \ldots, V_n \le v_n\}$, such as

$$\{V_2 \le v_2, V_1 \le v_1, \ldots, V_n \le v_n\} \text{ or } \{V_1 \le v_1, V_n \le v_n, \ldots, V_2 \le v_2\}.$$

Since there are $n!$ possible permutations, we conclude that

$$\mathbb{P}\left(V_{(1)} \le v_1, \ldots, V_{(n)} \le v_n\right) = n!\mathbb{P}\left(V_1 \le v_1, \ldots, V_n \le v_n\right) \overset{i.i.d.}{=} n! \prod_{i=1}^n v_i.$$

Compiled 09/12/2019 at 18:51

By formula

$$f_{V_{(1)},\dots,V_{(n)}}(v_1,\dots,v_n)=\frac{\partial^n}{\partial v_1\dots\partial v_n}\mathbb{P}\left(V_{(1)}\le v_1,\dots,V_{(n)}\le v_n\right)=n!I_A,$$

we derive that the joint probability density function is constant on A ($|A|=1/(n!)$) and vanishes outside A. Finally, in order to calculate the density function for the k-th order statistics we integrate over all other variables:

$$\begin{aligned}f_{V_{(k)}}(v_k)&=\int_{[0,1]^n} n!I_A\,dv_1\dots dv_n\\&=n!\int_0^{v_k}\int_{v_1}^{v_k}\cdots\int_{v_{k-2}}^{v_k} dv_{k-1}\dots dv_2\,dv_1\int_{v_k}^{v_{k+1}}\cdots\int_{v_k}^{v_n}\int_{v_k}^{1} dv_n\dots dv_{k+2}\,dv_{k+1}\\&=\frac{n!}{(k-1)!(n-k)!}v_k^{k-1}(1-v_k)^{n-k}=n\binom{n-1}{k-1}v_k^{k-1}(1-v_k)^{n-k}\end{aligned}\tag{6}$$

Problem 23. We consider the linear model with $n\ge p=2$, where

$$\mathbb{E}[y_j]=\beta_0,\quad j=1,\dots,n-1,\quad \mathbb{E}[y_n]=\beta_0+\beta_1.$$

a) Writing the model in the form $y=X\beta+\varepsilon$, find the least squares estimator of β as function of y_n and of $\bar{y}_0=(n-1)^{-1}\sum_{j_1}^{n-1}y_j$. Comment the form of this estimator.

b) Calculate the hat matrix for this model, verify that its trace is equal to p and find the adjusted values $\widehat{y}$.

c) Find the leverages h_{jj}, the standardised residuals and Cook's statistics. Comment on this.

Solution

a) From the fact that $\mathbb{E}[y]=X\beta$ we conclude that

$$X=\begin{pmatrix}1&0\\ \vdots&\vdots\\ 1&0\\ 1&1\end{pmatrix}\Longrightarrow X^\top X=\begin{pmatrix}n&1\\1&1\end{pmatrix}\Longrightarrow (X^\top X)^{-1}=\frac{1}{n-1}\begin{pmatrix}1&-1\\-1&n\end{pmatrix}.$$

The least squares estimator $\widehat{\beta}$ is computed as

$$\widehat{\beta}=(X^\top X)^{-1}X^\top y=\frac{1}{n-1}\begin{pmatrix}1&-1\\-1&n\end{pmatrix}\begin{pmatrix}\sum_{i=1}^{n-1}y_i+y_n\\ y_n\end{pmatrix}=\begin{pmatrix}\widetilde{y}\\ y_n-\widetilde{y}\end{pmatrix},$$

where $\widetilde{y}=\frac{1}{n-1}\sum_{i=1}^{n-1}y_i$.

b) By a direct calculation we have

$$H=X(X^\top X)^{-1}X^\top=\frac{1}{n-1}\begin{pmatrix}1&\dots&1&0\\ \vdots&\ddots&\vdots&\vdots\\ 1&\dots&1&0\\ 0&\dots&0&n-1\end{pmatrix},$$

and its trace is equal to $p=2$. The adjusted values are calculated as

$$\widehat{y}=Hy=\frac{1}{n-1}\begin{pmatrix}1&\dots&1&0\\ \vdots&\ddots&\vdots&\vdots\\ 1&\dots&1&0\\ 0&\dots&0&n-1\end{pmatrix}\begin{pmatrix}y_1\\ \vdots\\ y_{n-1}\\ y_n\end{pmatrix}=\begin{pmatrix}\widetilde{y}\\ \vdots\\ \widetilde{y}\\ y_n\end{pmatrix}.$$

c) The leverages are the diagonal values of the hat matrix. So,

$$h_{jj} = \begin{cases} \frac{1}{n-1} & \text{for } j = 1, \dots, n-1, \\ 1 & \text{for } j = n. \end{cases}$$

Next, we compute the standardised residuals. We start by computing the estimator for the variance σ^2 as

$$s^2 = \frac{1}{n-2} \sum_{j=1}^{n} (y_j - \widehat{y}_j)^2 = \frac{1}{n-2} \sum_{j=1}^{n-1} (y_j - \widetilde{y})^2,$$

and the standardised residuals are

$$r_i = \frac{y_i - \widehat{y}_i}{s\sqrt{1-h_{ii}}} = \begin{cases} \sqrt{\frac{(n-1)(n-2)}{\sum_{j=1}^{n-1}(y_j - \widetilde{y})^2}} (y_i - \widetilde{y}) & \text{for } i = 1, \dots, n-1, \\ 0 & \text{for } i = n. \end{cases}$$

The Cook's statistics are

$$C_i = \frac{r_i^2 h_{ii}}{p(1-h_{ii})} = \begin{cases} \frac{n-1}{n-2} \frac{(y_i - \widetilde{y})^2}{\sum_{j=1}^{n-1}(y_j - \widetilde{y})^2} & \text{for } i = 1, \dots, n-1, \\ 0 & \text{for } i = n. \end{cases}$$

The last observation, (x_n, y_n), is a leverage point, since we use the second parameter, β_1, just for fitting this observation, but it has no influence on all other data, since its Cook's statistic is zero.

Problem 24. (t-test)

Let $Y = X\beta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $X \in \mathbb{R}^{n \times p}$ of full column rank. Let us denote the t-statistic for the j-th parameter as

$$t = \frac{\widehat{\beta}_j - \beta_j}{\widehat{\mathrm{se}}(\widehat{\beta}_j)},$$

where $\mathrm{se}(\widehat{\beta}_j) = (\mathrm{var}(\widehat{\beta}_j))^{1/2}$ is the standard deviation of the estimator $\widehat{\beta}_j$ and $\widehat{\mathrm{se}}(\widehat{\beta}_j)$ is a suitable estimator of thereof. Show that $t \sim t_{n-p}$.

Solution

Firstly, recall that t_ν-distribution arises as

$$\frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_\nu}} \sqrt{\nu},$$

where the two random variables in the previous symbolic expression are independent.

Secondly, by the lemma on slide 98

$$\mathrm{var}(\widehat{\beta}_j) = \sigma^2 (X^\top X)^{-1}_{jj} := \sigma^2 w$$

and by the theorem on slide 74 we have $\frac{S^2}{\sigma^2}(n-p) \sim \chi^2_{n-p}$.

Combining the two facts, it is natural to estimate $\widehat{\sigma}^2 = S^2$ to get the estimator of the standard error. Note that by the theorem on slide 74 we also have the independence needed. Hence

$$t = \underbrace{\frac{\widehat{\beta}_j - \beta_j}{\sigma w}}_{\sim \mathcal{N}(0,1)} \; \underbrace{\frac{\sigma}{S\sqrt{n-p}}}_{\frac{1}{\sqrt{(n-p)\frac{S^2}{\sigma^2}}} \sim \frac{1}{\sqrt{\chi^2_{n-p}}}} \quad \sqrt{n-p} = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-p}}} \sqrt{n-p} \sim t_{n-p}.$$

Problem 25. When we adjust the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ to the cement data set (slide 90), R gives us the following table:

```
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   48.19363      3.91330    12.315   6.17e-07 ***
x1             1.69589      0.20458     8.290   1.66e-05 ***
x2             0.65691      0.04423    14.851   1.23e-07 ***
x3             0.25002      0.18471     1.354   0.209
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

a) Explain in details how we compute the values in the columns "`t value`" and "`Pr(>|t|)`". What is the meaning of these values? Comment the observed values.

b) Knowing that $\widehat{\text{corr}}(\widehat{\beta_1}, \widehat{\beta_2}) = -0.08911$, what is the p value for the null hypothesis $\beta_2 - \beta_3 = 0$? Try to find the value of the test statistics without using `R`. For a test with a threshold of 5%, can we reject the null hypothesis?

Solution

a) The column "`t value`" gives the statistics t for the hypothesis $\beta_i = 0$ defined by

$$T_i = \frac{\widehat{\beta}_i}{\sqrt{S^2 v_{ii}}} = \frac{\widehat{\beta}_i}{\widehat{\text{SE}}(\widehat{\beta}_i)},$$

where v_{ii} is the i-th diagonal element of the matrix $V = (X^\top X)^{-1}$. When the hypothesis $\beta_i = 0$ is true, we have that T_i follows a t Student lwa with with $n-p$ degrees of freedom. We will reject the null hypothesis $\beta_i = 0$ when the value of $|T_i|$ is large.

The column "`Pr(>|t|)`" gives the p-values for the bilateraul tests$t$ above. When we denote the observed value of T_i by τ_i, the p-value for the$i$-th test is given by

$$p_i = P(|T_i| > |\tau_i|) = 2(1 - t_{n-p}^{-1}(|\tau_i|)) = 2\,t_{n-p}^{-1}(-|\tau_i|)).$$

If $p_i < 0.05$, we reject the i-th hypothesis with a significance threshold of 5%.

For this example, with a significance threshold of 5%, we can reject the hypothesis $\beta_i = 0$ for $i = 0, 1, 2$, but not for $i = 3$.

b) In this case, the statistics t is given by

$$T = \frac{c^\top \widehat{\beta}}{\sqrt{S^2 c^\top (X^\top X)^{-1} c}}$$

for $c = [0, 0, 1, -1]^\top$. We know that

$$\begin{aligned} S^2 c^\top (X^\top X)^{-1} c &= \left(\widehat{\text{SE}}(\widehat{\beta}_2)\right)^2 + \left(\widehat{\text{SE}}(\widehat{\beta}_3)\right)^2 - 2\,\widehat{\text{corr}}(\widehat{\beta}_2, \widehat{\beta}_3)\,\widehat{\text{SE}}(\widehat{\beta}_2)\,\widehat{\text{SE}}(\widehat{\beta}_3) \\ &= 0.04423^2 + 0.18471^2 - 2 \cdot (-0.08911) \cdot 0.04423 \cdot 0.18471 = 0.03753. \end{aligned}$$

Hence

$$\tau = \frac{0.65691 - 0.25002}{\sqrt{0.03753}} = 2.10033$$

and we find the p-value

$$p = 2 \cdot t_9^{-1}(-2.10033) = 0.06508.$$

Thus, we do not reject the null hypothesis with a significance threshold of 5%.

Problem 26. Suppose the $n \times p$ full-rank design matrix X can be partitioned into two blocks as $[\,X_1\ X_2\,]$ and let $M_{X_1} := I_n - H_{X_1}$. Show that $H_X = H_{X_1} + H_{M_{X_1} X_2}$, where $H_{M_{X_1} X_2}$ is the projection on to the span of $M_{X_1} X_2$.

Solution
We need to show that $H_{X_1} + H_{M_{X_1}X_2}$ is an orthogonal projection matrix, i.e., it is idempotent, symmetric and it spans $\mathcal{S}(X)$. Note that $X_1^\top M_{X_1} X_2 = O$, so $H_{X_1} H_{M_{X_1}X_2} = O$ also. Since both $H_{M_{X_1}X_2}$ and H_{X_1} are orthogonal projection matrices, the first two statements are obvious.

It remains to show that any vector $z \in \mathcal{S}(X)$ is invariant under the action of $H_{X_1} + H_{M_{X_1}X_2}$ and that any vector orthogonal to this span is annihilated by $H_{X_1} + H_{M_{X_1}X_2}$. Since X is full rank, we can write $z = X\gamma = X_1\gamma_1 + X_2\gamma_2$ for some vector γ and subvectors γ_1 and γ_2. Then

$$\begin{aligned}(H_{X_1} + H_{M_{X_1}X_2})z &= (H_{X_1} + H_{M_{X_1}X_2})(X_1\gamma_1 + X_2\gamma_2)\\ &= H_{X_1}(X_1\gamma_1 + X_2\gamma_2) + H_{M_{X_1}X_2}(X_1\gamma_1 + X_2\gamma_2)\\ &= X_1\gamma_1 + H_{X_1}X_2\gamma_2 + M_{X_1}X_2\gamma_2\\ &= X_1\gamma_1 + X_2\gamma_2\end{aligned}$$

upon noting that

$$\begin{aligned}H_{M_{X_1}X_2}X_1 &= M_{X_1}X_2(X_2^\top M_{X_1}X_2)^{-1}X_2^\top M_{X_1}X_1 = O,\\ H_{M_{X_1}X_2}X_2 &= M_{X_1}X_2(X_2^\top M_{X_1}X_2)^{-1}X_2^\top M_{X_1}X_2 = M_{X_1}X_2.\end{aligned}$$

Take now $w \in \mathcal{S}^\perp(X)$. We have

$$\begin{aligned}(H_{X_1} + H_{M_{X_1}X_2})w &= H_{X_1}w + H_{M_{X_1}X_2}w\\ &= 0 + M_{X_1}X_2(X_2^\top M_{X_1}X_2)^{-1}X_2^\top M_{X_1}w\\ &= M_{X_1}X_2(X_2^\top M_{X_1}X_2)^{-1}X_2^\top(I - H_{X_1})w = 0.\end{aligned}$$

Indeed, $H_{X_1}w = 0$ because w is orthogonal to X, thus also orthogonal to X_1. At the same time, $X_2^\top w = 0$ by orthogonality. By uniqueness of projection matrices (Exercise 1.2), the result follows.

Problem 27. (Frisch–Waugh–Lovell theorem) Consider the linear regression $\boldsymbol{Y} = X_1\beta_1 + X_2\beta_2 + \varepsilon$ with $\mathsf{E}(\varepsilon) = 0_n$. Let y be the observed response and suppose the $n \times p$ full-rank design matrix X can be written as the partitioned matrix $[X_1\ X_2]$ with blocks X_1, an $n \times p_1$ matrix, and X_2, an $n \times p_2$ matrix. Let $\widehat{\beta}_1$ and $\widehat{\beta}_2$ be the ordinary least square (OLS) parameter estimates from running this regression. Suppose we run least squares on this model to obtain

$$\boldsymbol{y} = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + e, \tag{E1}$$

Define the orthogonal projection matrix H_X as usual and $H_{X_i} = X_i(X_i^\top X_i)^{-1}X_i^\top$ for $i = 1, 2$. Similarly, define the complementary projection matrices $M_{X_1} = I_n - H_{X_1}$ and $M_{X_2} = I_n - H_{X_2}$.

Prove the Frisch–Waugh–Lovell (FWL) theorem, i.e., show that the ordinary least square estimates $\widehat{\beta}_2$ and the residuals e from (E1) are identical to those obtained by running ordinary least squares on the regression

$$M_{X_1}\boldsymbol{y} = M_{X_1}X_2\beta_2 + \text{residuals}. \tag{E2}$$

Hint: starting from (E1) *assuming* $\widehat{\beta}_2$ *has been computed, pre-multiply both sides so as to obtain an expression in terms of* $\widehat{\beta}_2$ *only on the right-hand side and show the latter coincides with the least square estimate from* (E2).

Solution
The coefficient estimates of (E2) is

$$\widetilde{\beta}_2 = (X_2^\top M_{X_1}X_2)^{-1}X_2^\top M_{X_1}y. \tag{S1}$$

Let $\widehat{\beta}_1$ and $\widehat{\beta}_2$ denote the OLS estimates from running regression (E1). The orthogonal decomposition of y gives

$$y = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + M_X y. \tag{S2}$$

Premultiplying both sides of (S2) by $X_2^\top M_{X_1}$ yields

$$X_2^\top M_{X_1} y = X_2^\top M_{X_1} X_2 \widehat{\beta}_2 \tag{S3}$$

since $M_X M_{X_1} X_2 = M_X X_2 = O$. Solving (S3) gives back (S1), showing that $\widehat{\beta}_2 = \widetilde{\beta}_2$.

By premultiplying (S2) by M_{X_1}, we obtain instead

$$M_{X_1} y = M_{X_1} X_2 \widehat{\beta}_2 + M_X y \tag{S4}$$

since $M_{X_1} M_X = M_X$. The regressand in (S4) is the same as that of regression (E2). The first term, $M_{X_1} X_2 \widehat{\beta}_2$, must be the fitted value since $\widehat{\beta}_2$ is the OLS estimate of β_2. Thus, $M_X y$ must be the vector of residuals of (E2).

Deriving the expression for $\widehat{\beta}_2$ in the presence of multiple regressors involves tedious calculations with partitioned matrices. Use Frisch–Waugh-Lovell theorem when you have multiple regressors, but are only interested in a sub-vector of coefficient estimates such as $\widehat{\beta}_2$.

Problem 28. (t-test vs. F-test for model-submodel testing, requires the previous problem)

Consider the linear regression $y = X_1\beta_1 + x_2\beta_2 + \boldsymbol{\varepsilon}$ under the assumption that $X = (X_1^\top, x_2^\top)^\top$ is an $n \times p$ full-rank non-stochastic design matrix with x_2 an $n \times 1$ column vector and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$. We are interested in testing whether the parameter $\beta_2 = 0$: the Wald test t-statistic W and the Fisher test statistic F for this hypothesis are, respectively,

$$W = \frac{\widehat{\beta}_2}{\mathrm{se}(\widehat{\beta}_2)}, \qquad F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}/(n-p)},$$

where $\mathrm{se}(\widehat{\beta}_2) = \left[s^2 \mathsf{Var}\left(\widehat{\beta}_2\right)/\sigma^2\right]^{1/2}$. Under the null hypothesis $\mathscr{H}_0 : \beta_2 = 0$, $W \sim \mathcal{T}(n-p)$ and $F \sim \mathcal{F}(1, n-p)$. Show algebraically that $W^2 = F$.

Note that the two statistics lead to the same inference because the square of a $\mathcal{T}(n-p)$ distributed random variable has distribution $\mathcal{F}(1, n-p)$.

Solution

By the FWL theorem, we can write the arguments of W as

$$\widehat{\beta}_2 = (x_2^\top M_{X_1} x_2)^{-1} x_2^\top M_{X_1} y, \qquad \mathrm{se}(\widehat{\beta}_2) = \left[s^2 (x_2^\top M_{X_1} x_2)^{-1}\right]^{1/2}.$$

Clearly, $\mathrm{RSS}/(n-p) = s^2$ and thus it remains only to show that the numerator of F is

$$\mathrm{RSS}_0 - \mathrm{RSS} = \left[(x_2^\top M_{X_1} x_2)^{-1/2} x_2^\top M_{X_1} y\right]^2 = y^\top H_{M_{X_1} x_2} y.$$

First, we have

$$\mathrm{RSS}_0 - \mathrm{RSS} = \|M_{X_1} y\|^2 - \|M_X y\|^2 = \|(M_{X_1} - M_X) y\|^2.$$

Using an orthogonal decomposition, this expression can be further simplified to

$$\mathrm{RSS}_0 - \mathrm{RSS} = \|M_{X_1} H_X y\|^2 = \|M_{X_1}(H_{X_1} + H_{M_{X_1} x_2}) y\|^2 = \|H_{M_{X_1} x_2} y\|^2$$

because $H_{M_{X_1} x_2} \in \mathcal{M}^\perp(X_1)$. Noting that $\|H_{M_{X_1} x_2} y\|^2 = y^\top H_{M_{X_1} x_2} y$, completes the proof.

Problem 29. We consider the cement data with $n = 13$. The residuals sum of squares (RSS) for all the possible models (containing always the denoted variables and the intercept) are given below:

Model	RSS	Model	RSS	Model	RSS
- - - -	2715.8	1 2 - -	57.9	1 2 3 -	48.1
1 - - -	1265.7	1 - 3 -	1227.1	1 2 - 4	48.0
- 2 - -	906.3	1 - - 4	74.8	1 - 3 4	50.8
- - 3 -	1939.4	- 2 3 -	415.4	- 2 3 4	73.8
- - - 4	883.9	- 2 - 4	868.9		
		- - 3 4	175.7	1 2 3 4	47.9

Calculate the analysis of variance table (as in slide 146) when x_4, x_3, x_2 and x_1 are added to the model in this order, and test which term should be included in the model for the threshold $\alpha = 0.05$. Compare with slide 146.

Solution
Since the ordering of the variable is different, the number in the table will be different from the ones in slide 146. Namely:

	Df	Red Sum Sq	F value	p-value
x_4	1	2715.8-883.9 = 1831.9	306.3	10^-7
x_3	1	883.9 -175.7 = 708.2	118.4	10^-6
x_2	1	175.7 -73.8 = 101.9	17.04	0.003
x_1	1	73.8 -47.9 = 26	4.3	0.07
Residual	8	47.9		

For calculation of the F-values one calculates the numerator as the correspondent reduction in the sum of squares (third column of the table) divided by the degrees of freedom added (in this case only 1) and the denominator as the residual sum of squares divided by the residual degrees of freedom ($47.9/8 = 5.98$). Notice that the denominator stays always the same (which is quite irrelevant now in the computer era :). To calculate the p-values, use R: e.g. `p_val=1-df(F_val,1,8)`.

Adding variables in this (reverse) order would lead to a model with x_3 and x_4 (and with x_2 and arguably even x_1), while in the slide 161 variables x_1 and x_2 would be the only variables in the model.

Problem 30. (Orthogonal variables)
Let us consider the regression

$$y = X\beta + \varepsilon = (X_1, X_2)\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon,$$

where $X = (X_1, X_2)$, $\beta^\top = (\beta_1^\top, \beta_2^\top)$, X_1 is $n \times p_1$, X_2 is $n \times p_2$ (both injective) such that

$$X_1^\top X_2 = 0_{p_1 \times p_2}.$$

Let H_i be the hat matrix associated to X_i.

1. What is the geometrical interpretation of $X_1^\top X_2 = 0$?
2. Calculate H as a function of X_i and of H_i, then, calculate the products

$$H_1 H_2, H_2 H_1, H H_1, H_1 H.$$

   What do you notice, which is the geometrical interpretation?

3. Show that each of the following quantities are equal to Hy:
   (a) $H_1 y + H_2 y$;
   (b) $H_1 y + H_2 e_1$, with $e_1 = (I - H_1) y$;
   (c) $H_1 y + H e_1$.

Compiled 09/12/2019 at 18:51

4. Interpret these equalities in relation to the models

$$y = X\beta + \varepsilon \qquad (M)$$

and to its submodels

$$y = X_1\beta_1 + \varepsilon, \qquad (M_1)$$

$$y = X_2\beta_2 + \varepsilon. \qquad (M_2)$$

Solution

1. This means that all columns of X_1 are orthogonal to all columns of X_2. I.e., $\mathcal{M}(X_1) \perp \mathcal{M}(X_2)$.

2. We notice first that

$$X^\top X = \begin{pmatrix} X_1^\top X_1 & 0 \\ 0 & X_2^\top X_2 \end{pmatrix},$$

so

$$\begin{aligned} H &= (X_1, X_2)\begin{pmatrix} (X_1^\top X_1)^{-1} & 0 \\ 0 & (X_2^\top X_2)^{-1} \end{pmatrix}(X_1, X_2)^\top \\ &= X_1(X_1^\top X_1)^{-1}X_1^\top + X_2(X_2^\top X_2)^{-1}X_2^\top = H_1 + H_2. \end{aligned}$$

Then. since $X_1^\top X_2 = 0$, we have $H_1H_2 = 0$. Thus, $H_2H_1 = H_2^\top H_1^\top = (H_1H_2)^\top = 0$,

$$HH_1 = (H_1 + H_2)H_1 = H_1^2 = H_1$$

et $H_1H = H_1^\top H^\top = (HH_1)^\top = H_1^\top = H_1$.

Interpretation: $H_1H_2 = 0$ comes from the fact that the space of columns of X_1 and X_2 are orthogonal, thus if we project them on $\mathcal{M}(X_2)$ and then on $\mathcal{M}(X_1)$, we obtain the null vector. The interpretation for $H_2H_1 = 0$ is similar. $HH_1 = H_1$ comes from the fact that to project on $\mathcal{M}(X_1)$ and then on $\mathcal{M}(X)$ is equivalent to projecting only on $\mathcal{M}(X_1)$, because $\mathcal{M}(X_1)$ is a subspace of $\mathcal{M}(X)$. For the same reason, $H_1H = H_1$ because we project on $\mathcal{M}(X)$ and then on $\mathcal{M}(X_1)$, so it is the same as projecting on $\mathcal{M}(X_1)$. Intuitively, We notice that if $X_1^\top X_2 \neq 0$, We have $HH_1 = H_1 = H_1H$, but $H_1H_2 \neq 0$ and $H_2H_1 \neq 0$.

3. Using the fact that $Hy = (H_1 + H_2)y$,

   (a) trivial;

   (b) comes from $H_2H_1 = 0$;

   (c) comes from $H(I - H_1) = H - H_1 = H_2$.

4. The fitted values under (M) (with (y, X) as data) are equal to

   (a) the sum of the fitted values under (M_1) amd.d (M_2). (the model data (M_i) are (y, X_i))

   (b) the sum of the fitted values under (M_1) (given (y, X_1)) and of residuals of (M_1) fitted under (M_2) (the data are (e_1, X_2)).

   (c) the sum of the fitted values under (M_1) (given (y, X_1)) and of residuals of (M_1) fitted under (M) (the data are (e_1, X)).

Problem 31. (Orthogonal variables and ANOVA)

Let us consider the regression

$$y = X\beta + \varepsilon = (X_1, \ldots, X_k)\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \varepsilon$$

Compiled 09/12/2019 at 18:51

where X_i is $n \times p_i$, all the X_i are injective, and

$$i \neq j \implies X_i^\top X_j = 0.$$

Let H be the hat matrix associated to X, H_i the hat matrix associated to X_i and $\widehat{\beta} = (X^\top X)^{-1} X^\top y = (\widehat{\beta}_1^\top, \dots, \widehat{\beta}_k^\top)^\top$. We denote by δ_{ij} Kronecker's delta: $\delta_{ij} = 1$ if $i = j$, 0 otherwise. For a ordered set $L \subset \{1, \dots, k\}$ we define $X_L = (X_i : i \in L)$ and $\widehat{\beta}_L = (\widehat{\beta}_i^\top : i \in L)^\top$. For example, if $L = \{1, 2, 4\}$, $X_L = (X_1, X_2, X_4)$ and

$$\widehat{\beta}_L = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_4 \end{pmatrix}.$$

We define $RSS_L = \left\| y - H_L y \right\|^2$, where $H_L = X_L (X_L^\top X_L)^{-1} X_L^\top$.

1. Show that $H = H_1 + \cdots + H_k$ and that $H_L = \sum_{i \in L} H_i$.
2. Show that $H_i H_j = \delta_{ij} H_i$.
3. Show that $\widehat{\beta}_j = (X_j^\top X_j)^{-1} X_j^\top y$.
4. For $j \notin L$, calculate
$$RSS_L - RSS_{L \cup \{j\}},$$
and show that this expression does not depend on L.
5. Which is the interpretation of point 4. with respect to ANOVA?

Solution

1. since
$$(X^\top X)^{-1} = \begin{pmatrix} (X_1^\top X_1)^{-1} & 0 & \dots & 0 \\ 0 & (X_2^\top X_2)^{-1} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & (X_k^\top X_k)^{-1} \end{pmatrix}$$
and
$$(X_L^\top X_L)^{-1} = \operatorname{diag}((X_i^\top X_i)^{-1} : i \in L).$$
Son
$$H = X_1 (X_1^\top X_1)^{-1} X_1^\top + \cdots + X_k (X_k^\top X_k)^{-1} X_k^\top = H_1 + \cdots + H_k$$
and
$$H_L = \sum_{i \in L} X_i (X_i^\top X_i)^{-1} X_i^\top = \sum_{i \in L} H_i.$$
2. If $i = j$, $H_i H_j = H_i^2 = H_i$ and if $i \neq j$, $H_i H_j = X_i (X_i^\top X_i)^{-1} X_i^\top X_j (X_j^\top X_j)^{-1} X_j^\top = 0$ because $X_i^\top X_j = 0$.
3.
$$\widehat{\beta} = (X^\top X)^{-1} X^\top y = \begin{pmatrix} (X_1^\top X_1)^{-1} & 0 & \dots & 0 \\ 0 & (X_2^\top X_2)^{-1} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & (X_k^\top X_k)^{-1} \end{pmatrix} \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_k^\top \end{pmatrix} y = \begin{pmatrix} (X_1^\top X_1)^{-1} X_1^\top y \\ (X_2^\top X_2)^{-1} X_2^\top y \\ \vdots \\ (X_k^\top X_k)^{-1} X_k^\top y \end{pmatrix}.$$
4. First of all, we notice that
$$e_L := y - H_L y = y - \sum_{i \in L} H_i y$$

Compiled 09/12/2019 at 18:51

and

$$e_{L\cup\{j\}} := y - H_{L\cup\{j\}}y = y - \sum_{i\in L\cup\{j\}} H_i y.$$

Moreover,

$$\begin{aligned}(I - H_{L\cup\{j\}})e_L &= (I - H_{L\cup\{j\}})(I - H_L)y &(7)\\ &= (I - H_L - H_{L\cup\{j\}} + H_{L\cup\{j\}}H_L)y &(8)\\ &= (I - H_{L\cup\{j\}})y &(9)\\ &= e_{L\cup\{j\}}. &(10)\end{aligned}$$

Then $e_{L\cup\{j\}}$ is a orthogonal projection of e_L, so $e_L - e_{L\cup\{j\}} \perp e_{L\cup\{j\}}$ and

$$\|e_{L\cup\{j\}}\|^2 + \|e_L - e_{L\cup\{j\}}\|^2 = \|e_L\|^2.$$

So

$$RSS_L - RSS_{L\cup\{j\}} = \|e_L\|^2 - \|e_{L\cup\{j\}}\|^2 = \left\|e_L - e_{L\cup\{j\}}\right\|^2 = \left\|H_j y\right\|^2$$

is independent of L.

5. The interpretation with respect to ANOVA is that in this case, the addition of a variable X_j does not depend on the variables that we already have in the model (this is not the general case).

Problem 32. (Automatic model selection)
We consider the cement data. The residuals' sum of squares (RSS) and the Mallows' C_p for the model *containing the intercept* are the following:

Model	RSS	C_p	Model	RSS	C_p	Model	RSS	C_p
- - - -	2715.8	442.58	1 2 - -	57.9		1 2 3 -	48.1	
1 - - -	1265.7	202.39	1 - 3 -	1227.1	197.94	1 2 - 4	48.0	
- 2 - -	906.3		1 - - 4	74.8	5.49	1 - 3 4	50.8	
- - 3 -	1939.4	314.90	- 2 3 -	415.4	62.38	- 2 3 4	73.8	7.325
- - - 4	883.9	138.62	- 2 - 4	868.9	138.12			
			- - 3 4	175.7	22.34	1 2 3 4	47.9	5

1. Utilise the selection methods *forward selection* and *backward elimination* to chose some models for these data, including the significative variables at level 5%. Utilise the F-test

$$F = \frac{RSS(\widehat{\beta}_L) - RSS(\widehat{\beta}_{L\cup\{j\}})}{RSS(\widehat{\beta}_{\text{full}})/(13-5)}$$

to decide if the addition of the j-th variable is significative.

2. Another selection criterion is the Mallow's C_p:

$$C_p = \frac{SS_p}{s^2} + 2p - n.$$

Notice that here s^2 is the variance estimator in the complete model.

(a) How could we use this criterion? Calculate the missing C_p.

(b) Which is the model selected by this criterion using the *forward selection*, and then *backward elimination*? Among all the models considered, which one is the best, according to this criterion?

Compiled 09/12/2019 at 18:51

Solution

1. Here we will use the following test to add or not the j-th variable to the model $y = \beta_0 + \sum_{i\in L}\beta_i x_i$:

$$F = \frac{RSS(\widehat{\beta}_L) - RSS(\widehat{\beta}_{L\cup\{j\}})}{RSS(\widehat{\beta}_{\text{full}})/(13-5)},$$

where $\widehat{\beta}_{\text{full}}$ represents the estimator of β for the complete model. Since $RSS(\widehat{\beta}_L) - RSS(\widehat{\beta}_{L\cup\{j\}}) \sim \sigma^2\chi^2_1$ under the hypothesis $H_0 : \beta_j = 0$, and that $RSS(\widehat{\beta}_{\text{full}}) \sim \sigma^2\chi^2_{n-p}$ and it is independent of $RSS(\widehat{\beta}_L) - RSS(\widehat{\beta}_{L\cup\{j\}})$, $F \sim F_{1,8}$ under H_0. In particular, the distribution of F does not depend on the size of L. The critical value of this test at level 5% is 5.32.

Forward selection

- Initial model : $y = \beta_0 + \epsilon$
- Stage 1 : $y = \beta_0 + \beta_4 x_4 + \epsilon$, $F = \frac{2715.8-883.9}{47.9/(13-5)} = 305.95 > 5.32$.
- Stage 2 : $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \epsilon$, $F = 135.13 > 5.32$.
- Stage 3: $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $F = 4.47 < 5.32$.

Final model : $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \epsilon$.

Backward selection

- Initial model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$
- Stage 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \epsilon$, $F = \frac{48-47.9}{47.9/(13-5)} = 0.0167 < 5.32$.
- Stage 2 : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $F = 1.65 < 5.32$.
- Stage 3 : $y = \beta_0 + \beta_2 x_2 + \epsilon$, $F = 141.70 > 5.32$.

Final model : $y = \beta_0 + \beta_2 x_2 + \beta_1 x_1 + \epsilon$.

2. (a) Mallow's C_p work as AIC: we choose the model with the minimal C_p. Here's the table with all the C_p:

Model	RSS	C_p	Model	RSS	C_p	Model	RSS	C_p
- - - -	2715.8	442.58	1 2 - -	57.9	2.67	1 2 3 -	48.1	3.03
1 - - -	1265.7	202.39	1 - 3 -	1227.1	197.94	1 2 - 4	48.0	3.02
- 2 - -	906.3	142.37	1 - - 4	74.8	5.49	1 - 3 4	50.8	3.48
- - 3 -	1939.4	314.90	- 2 3 -	415.4	62.38	- 2 3 4	73.8	7.325
- - - 4	883.9	138.62	- 2 - 4	868.9	138.12			
			- - 3 4	175.7	22.34	1 2 3 4	47.9	5

   (b) With forward selection, we choose $y = \beta_0 + \sum_{i\in\{1,2,4\}}\beta_i x_i$, while the backward selection gives the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. This last model is the one with the smallest C_p among all others.

Problem 33. (AIC and Gaussian linear models)
Show that the AIC criterion for a Gaussian linear model, based on a response vector of size n, with p covariables and σ^2 unknown, can be written as :

$$\text{AIC} = n\log\widehat{\sigma}^2 + 2p + \text{const},$$

where $\widehat{\sigma}^2 = SS_p/n$ is the miximum likelihood estimator of σ^2

Solution
For the Gaussian linear models $y \sim N(X\beta, \sigma^2 I_n)$, the likelihood of (β, σ^2) is given by

$$L(\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}(y-X\beta)^\top(y-X\beta)\right).$$

Then, the log-likelihood is

$$l(\beta,\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-X\beta)^\top(y-X\beta).$$

Compiled 09/12/2019 at 18:51

We have that the maximum likelihood estimator of β and σ^2 are

$$\widehat{\beta} = (X^\top X)^{-1} X^\top y, \quad \widehat{\sigma}^2 = \frac{1}{n}(y - X\widehat{\beta})^\top (y - X\widehat{\beta}).$$

So, the maximum of log-likelihood is

$$l(\widehat{\beta}, \widehat{\sigma}^2) = -\frac{n}{2}\log(2\pi\widehat{\sigma}^2) - \frac{1}{2\widehat{\sigma}^2}\underbrace{(y - X\widehat{\beta})^\top (y - X\widehat{\beta})}_{=n\widehat{\sigma}^2} = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\widehat{\sigma}^2 - \frac{n}{2}.$$

From the AIC definition, we obtain that

$$\text{AIC} = -2l(\widehat{\beta}, \widehat{\sigma}^2) + 2p = n\log(2\pi) + n\log\widehat{\sigma}^2 + n + 2p = n\log\widehat{\sigma}^2 + 2p + \text{const.}$$

Problem 34. (Cross validation and number of regressions)

Let $y = X\beta + \epsilon$, $\widehat{\beta}$ denote the OLS estimator of β, X_{-j} denote the design matrix, which arises from X by dropping the k-th row x_k (which is understood for mathematical purposes as a column vector), and $\widehat{\beta}_{-j}$ be the estimator based on the data set without the k-th observation (symbolically, $y_{-k} = X_{-k}\beta_{-k} + \epsilon_{-k}$, again y_{-k} and ϵ_{-k} denote the vectors that arise from y and ϵ by dropping the k-th entry).

a) Using the Sherman-Morrison formula

$$\left(A + uv^\top\right)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u},$$

show that

$$(X_{-k}^\top X_{-k})^{-1} = \left(I + \frac{1}{1 - h_{kk}}\left(X^\top X\right)^{-1} x_k x_k^\top\right)\left(X^\top X\right)^{-1}.$$

b) Show that

$$X_{-k}^\top y = X^\top y - y_k x_k \quad \text{and} \quad x_k^\top (X^\top X)^{-1} X_{-k}^\top y = (1 - h_{kk})y_k - e_k,$$

to conclude that

$$\widehat{\beta}_{-k} = \widehat{\beta} - \frac{e_k\left(X^\top X\right)^{-1} x_k}{1 - h_{kk}}.$$

c) Use the previous formula to show that the cross-validation criterion

$$\text{CV} = \sum_{j=1}^{n} (y_j - x_j^\top \widehat{\beta}_{-j})^2. \tag{11}$$

can be written as

$$\text{CV} = \sum_{j=1}^{n} \frac{(y_j - x_j^\top \widehat{\beta})^2}{(1 - h_{jj})^2}. \tag{12}$$

What is the advantage of using (12) instead of (11)?

Solution

a) First note that $X_{-k}^\top X_{-k} = X^\top X - x_k x_k^\top$. Denoting $C = X^\top X$ and using Sherman-Morrison we have

$$\begin{aligned}
(X_{-k}^\top X_{-k}^\top)^{-1} &= (C - x_k x_k^\top)^{-1} \\
&= C^{-1} + \frac{C^{-1}x_k x_k^\top C^{-1}}{1 - x_k^\top C^{-1} x_k} \\
&= \left(I + \frac{C^{-1}x_k x_k^\top}{1 - h_{kk}}\right) C^{-1}, \\
&= \left(I + \frac{(X^\top X)^{-1}x_k x_k^\top}{1 - h_{kk}}\right) (X^\top X)^{-1},
\end{aligned}$$

b) Similarly to the previous part, we can calculate

$$X^\top y = (x_1, \dots, x_n) y = \sum_{i=1}^{n} x_i y_i = X_{-k}^\top y + x_k y_k$$

and

$$\begin{aligned} x_k^\top (X^\top X)^{-1} X_{-k}^\top y &= x_k^\top (X^\top X)^{-1} (X^\top y - x_k y_k) \\ &= \widehat{y}_k - h_{kk} y_k \\ &= y_k - e_k - h_{kk} y_k \\ &= (1 - h_{kk}) y_k - e_k. \end{aligned}$$

Now we can calculate

$$\begin{aligned} \widehat{\beta}_{-k} &= (X_{-k}^\top X_{-k})^{-1} X_{-k}^\top y_{-k} \\ &= (X_{-k}^\top X_{-k})^{-1} X_{-k}^\top y \\ &= \left(I + \frac{(X^\top X)^{-1} x_k x_k^\top}{1 - h_{kk}} \right) (X^\top X)^{-1} X_{-k}^\top y \\ &= (X^\top X)^{-1} (X^\top y - y_k x_k) + (1 - h_{kk})^{-1} (X^\top X)^{-1} x_k x_k^\top (X^\top X)^{-1} X_{-k}^\top y \\ &= \widehat{\beta} - (X^\top X)^{-1} x_k y_k + (1 - h_{kk})^{-1} (X^\top X)^{-1} x_k [(1 - h_{kk}) y_k - e_k] \\ &= \widehat{\beta} - (1 - h_{kk})^{-1} e_k (X^\top X)^{-1} x_k. \end{aligned}$$

c) Now we know that

$$\widehat{\beta}_{-j} = \widehat{\beta} - \frac{(y_j - \widehat{y}_j) \left(X^\top X \right)^{-1} x_j}{1 - h_{jj}}.$$

So, we have

$$\begin{aligned} x_j^T \widehat{\beta}_{-j} &= x_j^T \widehat{\beta} - (1 - h_{jj})^{-1} x_j^T (X^T X)^{-1} x_j (y_j - \widehat{y}_j) \\ &= \widehat{y}_j - \frac{h_{jj}}{1 - h_{jj}} (y_j - \widehat{y}_j) \\ &= \widehat{y}_j + \left(1 - \frac{1}{1 - h_{jj}} \right) (y_j - \widehat{y}_j) \\ &= \widehat{y}_j + y_j - \widehat{y}_j - \frac{1}{1 - h_{jj}} (y_j - \widehat{y}_j) \end{aligned}$$

from which

$$y_j - x_j^T \widehat{\beta}_{-j} = \frac{1}{1 - h_{jj}} (y_j - \widehat{y}_j)$$

follows.

If we use formula (11) we have to estimate all the β_{-j}, $j = 1, \dots, n$, and then proceed to n adjustments. On the other hand, if we use formula (12) only the adjustment of the model with the complete data is required. This makes it feasible to actually perform "leave-one-out" cross-validation for a linear model.

Problem 35. Let us suppose that $y = \mu + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and that we adjusted to y a linear model with the full rank design matrix $X_{n \times p}$, $n \geq p$, and the corresponding hat matrix H. Let D be the diagonal matrix with elements $1 - h_{11}, \dots, 1 - h_{nn}$. Using the previous exercise, show that

$$\mathbb{E}[CV] = \mu^\top (I - H) D^{-2} (I - H) \mu + \sigma^2 tr(D^{-1}),$$

and deduce that if μ belongs to the space generated by the columns of X, then $\mathbb{E}[CV] \approx (n+p)\sigma^2$.

Solution
The cross validation CV is defined equivalently as (see previous exercise) $CV = \sum_{j=1}^{n} \frac{(y_j - x_j^\top \widehat{\beta})^2}{(1-h_{jj})^2}$ or, in vectorial notation,

$$\begin{aligned} CV &= (y - X\widehat{\beta})^\top D^{-2}(y - X\widehat{\beta}) \\ &= (y - Hy)^\top D^{-2}(y - Hy) \\ &= [D^{-1}(I-H)y]^\top [D^{-1}(I-H)y] \end{aligned}$$

From the rule $\mathbb{E}[Z^\top Z] = \mathbb{E}[Z]^\top \mathbb{E}[Z] + \mathrm{tr}(\mathrm{cov}(Z))$ for a random vector z we derive

$$\begin{aligned} \mathbb{E}[CV] &= \mathbb{E}[D^{-1}(I-H)y]^\top \mathbb{E}[D^{-1}(I-H)y] + \mathrm{tr}(\mathrm{cov}(D^{-1}(I-H)y)) \\ &= \mathbb{E}[y]^\top (I-H)^\top D^{-2}(I-H)\mathbb{E}[y] + \mathrm{tr}(D^{-1}(I-H)\mathrm{cov}(y)(I-H)^\top D^{-1}) \\ &= \mathbb{E}[y]^\top (I-H)^\top D^{-2}(I-H)\mathbb{E}[y] + \sigma^2 \mathrm{tr}(D^{-1}(I-H)(I-H)^\top D^{-1}) \\ &= \mathbb{E}[y]^\top (I-H)^\top D^{-2}(I-H)\mathbb{E}[y] + \sigma^2 \mathrm{tr}(D^{-1}(I-H)D^{-1}), \end{aligned}$$

since $I-H$ is a projection matrix (symmetric and idempotent). We prove that $\mathrm{tr}(D^{-1}(I-H)D^{-1}) = \mathrm{tr}(D^{-1})$ as last step (δ_{ij} is the Kronecker delta):

$$\begin{aligned} \mathrm{tr}(D^{-1}(I-H)D^{-1}) &= \sum_{i=1}^{n} (1-h_{ii})^{-1}\delta_{ij}(\delta_{jk} - h_{jk})(1-h_{kk})^{-1}\delta_{ki} \\ &= \sum_{i=1}^{n} (1-h_{ii})^{-1}(\delta_{ii} - h_{ii})(1-h_{ii}) \\ &= \sum_{i=1}^{n} (1-h_{ii})^{-1} = \mathrm{tr}(D^{-1}). \end{aligned}$$

When μ belongs to the column space of X, $\mathcal{M}(X)$, then $(I-H)\mu = 0$, as $I-H$ is the projection matrix on the orthogonal subspace $\mathcal{M}(X)^\perp$. So, in this case

$$\mathbb{E}[CV] = \sigma^2 tr(D^{-1}) = \sigma^2 \sum_{i=1}^{n} \frac{1}{1-h_{ii}} \approx \sigma^2 \sum_{i=1}^{n} (1+h_{ii}) = \sigma^2(n + tr(H)) = \sigma^2(n+p),$$

since h_{ii} are usually small (high leverage pose issues for the model) and function $f(x) = 1/(1-x)$ has derivative at 0 equal to one, so from the Taylor expansion at 0 it is $f(x) \approx 1 + x$.

Problem 36. (Model selection in `R`)

a) Use the criteria *backward stepwise* and *forward stepwise* to choose a model for the data"Supervisor Performance" (`SPD`) from `R` package `RSADBE`

   Which model has the best AIC value?

b) Using the package `leaps`, find the model with the best BIC value among all submodels.

Solution
a)
```
library(RSADBE)
data(SPD)

m1 <- lm(Y ~ ., data = SPD)
m.backward <- step(m1, direction = "backward")

m0 <- lm(Y ~ 1, data = SPD)
my.scope <- formula(SPD)
```


Compiled 09/12/2019 at 18:51

```
m.forward <- step(m0, scope = my.scope, direction = "forward", data = SPD)
```

The forward/backward stepwise give the model `Y ~ X1 + X3` with AIC of 118.00.

b)
```
install.packages("leaps")
library(leaps)
library(car)

leaps <- regsubsets(formula(perf), data = perf)
plot(leaps)
subsets(leaps)
```

The model with the best BIC value is `Y ~ X1` with BIC of -27.50.

Problem 37. (Ridge regression)

Let $X = [1_n\ Z]$ be an $n \times p$ design matrix with centered inputs Z, meaning that $Z^\top 1_n = 0_{p-1}$. Consider the model $y = 1_n\alpha + Z\gamma + \boldsymbol{\varepsilon}$, where $\mathsf{E}(\boldsymbol{\varepsilon}) = 0_n$ and $\mathsf{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$.

a) Show that the fitted value of the ridge regression are

$$\widehat{y}_{\text{ridge}} = \overline{y}1_n + \sum_{j=1}^{p-1} \frac{d_j^2}{d_j^2 + \lambda} u_j u_j^\top y,$$

where u_j are the left singular column vectors of Z and d_j the elements of a diagonal matrix to be determined. Discuss what happens to $\widehat{y}_{\text{ridge}}$ when some of the $\{d_j^2\}_{j=1}^{p-1}$ are close to zero.

b) What happens to the ridge estimates if the columns of Z are orthogonal? Explain why it is preferable to standardize the columns of Z so they have approximately unit variance.

c) Show that $\lambda \mapsto \left\|\widehat{\gamma}_{\text{ridge}}\right\|_2^2$ is a decreasing function.

Solution

a) Using the SVD decomposition $Z = UDV^\top$, we have

$$Z\widehat{\gamma}_{\text{ridge}} = Z(Z^\top Z + \lambda I_{p-1})^{-1} Z^\top y = UD(D^2 + \lambda I_{p-1})^{-1} DU^\top y = \sum_{j=1}^{p-1} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^\top y.$$

The values d_j^2 are the square of the singular values of Z, so they correspond to the eigenvalues of $Z^\top Z$. The coefficients associated to the basis vectors u_j with smallest eigenvalues $\omega_j = d_j^2$ get shrunk the most towards zero.

b) Since $Z^\top Z = I_{p-1}$, the shrinkage is uniform over all variables. The OLS are invariant to scaling of the variables (in that a scaling of a column $x_j \mapsto kx_j$ leads to the solution $\widehat{\gamma}_j \mapsto \widehat{\gamma}_j / k$. This is **not** the case for penalized methods, as it implies implying different amount of shrinkage to the covariates. To see this, note that the minimizer of the objective function $\|y - 1_n\beta_0 + \Lambda Z\gamma\|^2 + \lambda\|\gamma\|_1^2$ changes. Imposing a common standard deviation ensures that the penalty is consistent and is automatically done by any good R package.

c) Let $Z = UDV^\top$ be a singular value decomposition of Z with D a $(p-1) \times (p-1)$ diagonal matrix and U and $n \times (p-1)$ orthonormal matrix. This gives us the eigendecomposition of $Z^\top Z = VD^2V^\top$ with $D^2 = \Lambda$ the diagonal matrix with the ordered eigenvalues ω_j^2. We can write the coefficient $\widehat{\gamma}$ as

$$\widehat{\gamma}_{\text{ridge}} = V(D^2 + \lambda I_{p-1})^{-1} DU^\top y = \sum_{j=1}^{p-1} \frac{\omega_j}{\omega_j^2 + \lambda} (u_j^\top y) v_j.$$

Since V is orthogonal,

$$\left\|\widehat{\gamma}_{\text{ridge}}\right\|_2^2 = y^\top UD(D^2+\lambda I_{p-1})^{-1}V^\top V(D^2+\lambda I_{p-1})^{-1}DU^\top y = \|\Omega U^\top y\|_2^2 = \sum_{j=1}^{p-1}\left(\frac{\omega_j}{\omega_j^2+\lambda}\right)^2 (u_j^\top y)^2,$$

where the diagonal matrix $\Omega := (D^2+\lambda I_{p-1})^{-1}D$. The expression is decreasing in λ.

Problem 38. Let $\lambda^* = 2\max_{1\le j\le q}|Z_j^\top y|$. Show that

$$\begin{cases} \lambda > \lambda^* \implies \widehat{\gamma}_{\text{lasso}} = 0, \\ \lambda < \lambda^* \implies \widehat{\gamma}_{\text{lasso}} \neq 0. \end{cases}$$

Hint: Use the convexity for the first part.

Solution

Let $\widetilde{y} = y - \bar{y}\mathbb{1}$ is the centered response. Then $\widehat{\gamma} = \widehat{\gamma}_{\text{lasso}}$ minimizes function f defined as

$$f(\gamma) = g(\gamma) + \lambda\|\gamma\|_1 \quad \text{with} \quad g(\gamma) = \sum_{i=1}^{n}\left(\widetilde{y}_i - \sum_{j=1}^{q} Z_{ij}\gamma_j\right)^2.$$

We will study what happens with the two parts of f close to 0. For g, this will be done via derivative, while the non-differentiable term $\|\gamma\|_1$ we will be inspected directly (another approach would be to use the sub-gradient, an optimization-theory notion generalizing the concept of a derivative).

The partial derivatives of g at 0 are

$$\frac{\partial g}{\partial \gamma_j}(0) = -\sum_{i=1}^{n} 2\Big(\widetilde{y}_i - \sum_{j=1}^{q} Z_{ij}0\Big)Z_{ij} = -2Z_j^\top \widetilde{y} = -2Z_j^\top y, \quad j = 1,\dots,q,$$

where the last equality comes from the fact that $Z^\top\mathbb{1} = 0$.

For the case $\lambda < \lambda^*$, we will show that there exist v such that $f(v) < f(0)$. Let j be the coordinate for which $\lambda < 2|Z_j^\top y|$, and let e_j denote the j-th vector of the standard basis (i.e. zero but 1 in the j-th coordinate). For t small we have

$$f(te_j) = g(te_j) + \lambda\|te_j\| = g(te_j) + \lambda|t| = g(0) + t\big[-2Z_j^\top y + \lambda\text{sign}(t) + o(1)\big].$$

If $Z_j^\top y > 0$, $f(te_j) < g(0) = f(0)$ for $t > 0$ small enough. If $Z_j^\top y < 0$, the same is true for $t < 0$ small enough. Hence 0 is not the minimizer of f.

Now let $\lambda > \lambda^*$. We can estimate, using the Taylor expansion for g at 0, that for any v:

$$f(v) = g(0) + [\nabla g(0)]^\top v + o(v) + \lambda\|v\|_1 \ge g(0) + \big(\lambda - \underbrace{\|\nabla g(0)\|_\infty}_{\lambda^*}\big)\|v\|_1 + o(v).$$

Recall that $o(v)/\|v\|_1 \to 0$ for $v \to 0$. Hence, since $\lambda > \lambda^*$, 0 must be a strict local minimum of f. Since f is convex, 0 must be the only minimum.

Problem 39. Unlike the ridge regression, lasso solution is not always unique. However, the adjusted values are unique: let $\widehat{\beta}_1$ and $\widehat{\beta}_2$ be two lasso solutions (for the same smoothing parameters λ).

a) Show that $X\widehat{\beta}_1 = X\widehat{\beta}_2$, using convexity.

b) Show that, if $\lambda > 0$, then $\|\widehat{\beta}_1\|_1 = \|\widehat{\beta}_2\|_1$.

Solution

Since both the estimators estimate the intercept the same (as the mean), so we can only concentrate on Z and γ estimates, denoted as $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$.

a) Assume that $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ both give an optimal objective value, henceforth denoted as α. Note first that $\|Y - Z\gamma\|_2^2$ is a strictly convex function of $Z\gamma$ and hence for $t \in (0,1)$ we have

$$\|Y - tZ\widehat{\gamma}_1 - (1-t)Z\widehat{\gamma}_2| \leq t\|Y - Z\widehat{\gamma}_1\|_2^2 + (1-t)\|Y - Z\widehat{\gamma}_1\|_2^2. \tag{13}$$

By strict convexity, we know that the equality appears only if $Z\widehat{\gamma}_1 = Z\widehat{\gamma}_2$. Using optimality of $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ and convexity of the L^1-norm we obtain

$$\begin{aligned}
\alpha &\leq \|Y - tZ\widehat{\gamma}_1 - (1-t)Z\widehat{\gamma}_2\|_2^2 + \lambda\|t\widehat{\gamma}_1 + (1-t)\widehat{\gamma}_2\|_1 \\
&\leq t\|Y - Z\widehat{\gamma}_1\|_2^2 + (1-t)\|Y - Z\widehat{\gamma}_2\|_2^2 + \lambda t\|\widehat{\gamma}_1\|_1 + \lambda(1-t)\|\widehat{\gamma}_2\|_1 \\
&= t\Big(\|Y - Z\widehat{\gamma}_1\|_2^2 + \lambda\|\widehat{\gamma}_1\|_1\Big) + (1-t)\Big(\|Y - Z\widehat{\gamma}_2\|_2^2 + \lambda\|\widehat{\gamma}_2\|_1\Big) \\
&= t\alpha + (1-t)\alpha = \alpha.
\end{aligned}$$

Hence equalities must be preserved in the previous chain, hence (13) also must be equality and hence we get what we wanted.

b) This is evident from the previous part, because in the last inequality we used two upper estimates. If the inequality should be equality, both of the upper estimates must be sharp. From the sharpness of the first one we deduced part a), from the second one we can deduce part b), provided $\lambda > 0$.

Problem 40. (Median regression)
Let $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$. Note that the median of a random variable Y is defined as

$$\operatorname{med}(Y) = \operatorname*{arg\,min}_{c \in \mathbb{R}} \mathsf{E}|Y - c|.$$

Let $X_i = (1, x_i)^\top$ and

$$\widehat{\beta} = \operatorname*{arg\,min}_{\beta} \sum (Y_i - \beta^\top X_i)^2, \qquad \widetilde{\beta} = \operatorname*{arg\,min}_{\beta} \sum |Y_i - \beta^\top X_i|$$

1. Show that $\mathsf{E}|Y - \beta^\top X|$ is minimized for $\beta^\top X = \operatorname{med}(Y)$ and conclude why $\widetilde{\beta}$ is sometimes called the "median regression estimate". *(Note that X denotes here a generic vector of regressors just as Y denotes a generic random variable from which Y_i's are sampled.)*

2. Compare what are the estimators $\widehat{\beta}$ and $\widetilde{\beta}$ actually estimating in the cases of $\epsilon \sim N(0,1)$ and $\epsilon_i \sim Exp(1)$.

Solution

1. This is clear from how median is defined. $\widetilde{\beta}$ is modeling median of the response variable in the same way as $\widehat{\beta}$ is modeling the expectation.

2. Since both the median and expectation of a standard gaussian is zero, in the gaussian case the two estimators are estimating the same:
$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 x_i = \operatorname{med}(Y_i).$$
In the second case, $\operatorname{med}(Exp(1)) \neq \mathsf{E}\big(Exp(1)\big)$ and both of them are non-zero, hence $\widehat{\beta}_0$ is estimating a different constant than $\widetilde{\beta}_0$, and neither is really estimating β_0. β_1 is being estimated the same, since both the median and the expectation are linear. Hence the effect of the covariate on the response is the same in both cases.

Problem 41. (CV and ridge regression)

1. Explain how to use cross-validation to choose the tuning parameter λ for ridge regression.

2. In the spirit of Problem 34, show that leave-one-out CV is computationally tractable in case of ridge regression.

Compiled 09/12/2019 at 18:51

Solution

1. To perform leave-one-out CV, we first have to choose values of λ on some grid of size K of possible values we wish to investigate, i.e. $\lambda = \lambda_1, \dots, \lambda_K$. Since $\lambda \in (0, \infty)$, we may use transformation of $(0,\infty)$ to some compact interval, e.g. $\lambda \mapsto 1/(1+\lambda)$ maps $(0,\infty) \mapsto (0,1)$, then we take a regular grid $\{t_1, \dots, t_K\} \in (0,1)$ and take $\lambda_k = (1-t_k)/t_k$.

   Once we have a set of λ's, whose suitability we wish to compare, we calculate the leave-one-out cross-validation criterion for all of them:

   $$CV(\lambda_k) = \sum_{j=1}^{n} \left(y_j - x_j^\top \widehat{\beta}_{-j}(\lambda_k) \right)^2, \quad k = 1, \dots, K.$$

   where $\widehat{\beta}_{-j}(\lambda_k)$ is the ridge regression estimator with the tuning parameter $\lambda = \lambda_j$ fitted without the j-th observation. Finally, we choose the best λ (among the considered ones) as $\widehat{\lambda} := \operatorname{arg\,min} CV(\lambda_k)$.

2. We want to show that

   $$CV(\lambda_k) = \sum_{j=1}^{n} \frac{\left(y_j - x_j^\top \widehat{\beta}(\lambda_k) \right)^2}{1 - s_{jj}(\lambda_k)}, \quad k = 1, \dots, K,$$

   where $S(\lambda_k) = X(X^\top X + \lambda_k I) X^\top$ and $s_{jj}(\lambda_k)$ is the (j,j)-th entry of $S(\lambda_k)$. Knowing this allows us to save a lot of computations, because it is clear from the formula that to perform leave-one-out CV, one only has to fit a single ridge regression to every considered value of λ (the same as for the regular regression).

   To prove this assertion, we go through the proof in the solution to Problem 34, taking replacing $X^\top X$ by $X^\top X + \lambda I$, leading to h_{jj} replaced by s_{jj} everywhere.

   Note: A similar result can be shown more generally for any model, which calculates the fitted values $\widehat{y}$ as a linear transformation of the observed values y. This is the case for both the ordinary linear regression and the ridge regression, but not for the lasso. In the case of lasso, leave-one-out CV is thus computationally intractable, hence one rather uses the so-called K-fold CV (mostly with $K = 5$ or $K = 10$).

Problem 42. (Generalized least squares)
Consider the linear model $Y = X\beta + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y}$ is an $n \times 1$ vector of responses, X is an $n \times p$ full-rank non-stochastic design matrix and the error vector $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0_n, \Sigma)$ for $\Sigma \neq \sigma^2 I_n$ a *known* positive definite covariance matrix. Let y be the observed response vector.

1. Show that the maximum likelihood estimator (MLE) of β is the vector that minimizes

   $$(y - X\beta)^\top \Sigma^{-1} (y - X\beta).$$

2. Show that the maximum likelihood estimator of β, known as generalized least squares estimator (GLS), is of the form

   $$\widehat{\beta}_{\mathrm{GLS}} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} y.$$

3. Derive the distribution of $\widehat{\beta}_{\mathrm{GLS}}$.

4. Show that the ordinary least squares (OLS) estimator $\widehat{\beta}$ is an unbiased estimator of β, but is not the best linear unbiased estimator (BLUE) of β. State carefully any result you use.

Solution

1. The maximum likelihood estimator of β is

   $$\operatorname*{arg\,max}_{\beta} L(y; X) = \operatorname*{arg\,max}_{\beta} \frac{|\Sigma|^{-1/2}}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2} (y - X\beta)^\top \Sigma^{-1} (y - X\beta) \right\}.$$

   and thus finding the MLE amounts to minimization of $(y - X\beta)^\top \Sigma^{-1} (y - X\beta)$.

Compiled 09/12/2019 at 18:51

2. Since $\Sigma = U\Lambda U^\top$ is positive definite, its inverse Σ^{-1} is well-defined and positive definite and by the spectral theorem admits a square root $\Sigma^{-1/2} = U\Lambda^{-1/2}U^\top$.

   One can rewrite the regression as the classical linear model setting by premultiplying by $\Sigma^{1/2}$. The normal equation can also be derived using vector calculus, by differentiating $(y - X\beta)^\top\Sigma^{-1}(y - X\beta)$ with respect to β and setting the derivative to zero. The normal equations are

$$X^\top\Sigma^{-1}X\beta = X^\top\Sigma^{-1}y$$

   and since $X^\top\Sigma^{-1}X$ is a quadratic form and Σ is positive definite, the inverse is well-defined. Differentiating twice gives $2X^\top\Sigma^{-1}X$ and since the Hessian is positive, $\widehat{\beta}_{\text{GLS}}$ minimizes the distance and is therefore the maximum likelihood estimator of β.

3. By the transformation property, the estimator $\widehat{\beta}_{\text{GLS}}$ is Gaussian because $\boldsymbol{\varepsilon}$ is also Gaussian. Its mean and variance are

$$\mathsf{E}\left(\widehat{\beta}_{\text{GLS}}\right) = (X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}\mathsf{E}\left(Y\right) = (X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}X\beta = \beta$$

   and

$$\begin{aligned}\mathsf{Var}\left(\widehat{\beta}_{\text{GLS}}\right) &= (X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}\mathsf{Var}\left(Y\right)\Sigma^{-1}X(X^\top\Sigma^{-1}X)^{-1}\\ &= (X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}\Sigma\Sigma^{-1}X(X^\top\Sigma^{-1}X)^{-1}\\ &= (X^\top\Sigma^{-1}X)^{-1}.\end{aligned}$$

4. First, the ordinary least square (OLS) estimator is unbiased,

$$\mathsf{E}\left(\widehat{\beta}_{\text{OLS}}\right) = (X^\top X)^{-1}X^\top\mathsf{E}\left(Y\right) = \beta.$$

   Let $Y^* = \Sigma^{-1/2}Y$. Then, the linear model with $Y^* = \Sigma^{-1/2}X\beta + \boldsymbol{\varepsilon}^*$ satisfies the hypothesis of the Gauss–Markov theorem with $\boldsymbol{\varepsilon}^* \sim \mathcal{N}_n(0_n, I_n)$ and the OLS estimator of this regression is BLUE. Since we premultiply by the matrix $\Sigma^{-1/2}$, the design matrix becomes $\Sigma^{-1/2}X$ and so the BLUE estimator is $\widehat{\beta}_{\text{GLS}}$.

   Alternatively, proceed as in the proof of Gauss–Markov theorem to show that $\widehat{\beta}_{\text{GLS}}$ is BLUE.

   Let $\widetilde{\beta}$ be any linear unbiased estimator of β, necessarily of the form AY with $AX = I_n$. Write $\mathsf{Var}\left(\widetilde{\beta}\right) = A\Sigma A^\top$ and the difference between the variance of the estimators as

$$\begin{aligned}\mathsf{Var}\left(\widetilde{\beta}\right) - \mathsf{Var}\left(\widehat{\beta}_{\text{GLS}}\right) &= A\Sigma A^\top - (X^\top\Sigma^{-1}X)^{-1}\\ &= A\left\{\Sigma - X(X^\top\Sigma^{-1}X)^{-1}X^\top\right\}A^\top\\ &= A\Sigma^{1/2}\left\{I_n - \Sigma^{-1/2}X(X^\top\Sigma^{-1/2}\Sigma^{-1/2}X)^{-1}X^\top\Sigma^{-1/2}\right\}\Sigma^{1/2}A^\top\\ &= A\Sigma^{1/2}M_{\Sigma^{-1/2}X}\Sigma^{1/2}A^\top.\end{aligned}$$

   Since $M_{\Sigma^{-1/2}X}$ is a projection matrix, it is idempotent and the difference $\mathsf{Var}\left(\widetilde{\beta}\right) - \mathsf{Var}\left(\widehat{\beta}_{\text{GLS}}\right)$ is a quadratic form and hence positive semi-definite. Since $\widehat{\beta}_{\text{OLS}}$ is a linear unbiased estimator (with $A = (X^\top X)^{-1}X^\top$), it is not the BLUE in this particular example.

Problem 43. Consider the linear model $y = X\beta + \varepsilon$, with $\varepsilon_j \overset{iid}{\sim} g(\cdot)$; suppose that $\mathbb{E}(\varepsilon_j) = 0$ and $\text{var}(\varepsilon_j) = \sigma^2 < \infty$ is known. Suppose that the MLE of β is regular, with

$$i_g = \int -\frac{\partial^2 \log g(u)}{\partial u^2} g(u)\,du = \int \left\{\frac{\partial \log g(u)}{\partial u}\right\}^2 g(u)\,du.$$

1. Show that the asymptotic relative efficiency (ARE) of the leas squares estimator of β relative to MLE of β

is

$$\frac{1}{\sigma^2 i_g}.$$

2. What is it reduced to if g is the gaussian density?

3. What about if g is the density of the Laplace distribution?

Solution

Note: In this exercise we use $x_j^\top$ to denote the j-th row of X, i.e. $X = (x_1|x_2|\ldots|x_n)^\top$. Also recall that the MLE of θ is asymptotically Gaussian with the covariance matrix being the inverse of the Fisher information matrix.

The density of y_j is $g(y_j - x_j^\top\beta)$ where g is the density of ϵ_j. Thus we have

$$\ell(\beta) = \sum_{j=1}^{n} \log g(y_j - x_j^\top\beta), \qquad \beta \in \mathbb{R}^p.$$

Denoting $h_j(\beta) := \log g(y_j - x_j^\top\beta)$, we have from the chain rule $\frac{\partial(Af(\beta))}{\partial\beta} = \frac{\partial(f(\beta))}{\partial\beta}A^\top$ (with a vector function f) that

$$\frac{\partial h_j}{\partial\beta} = -x_j \left.\frac{d\log g(u)}{du}\right|_{u=y_j-x_j^\top\beta}$$

and hence

$$\frac{\partial^2 h_j}{\partial\beta^2} = \frac{\partial}{\partial\beta}\frac{\partial h_j}{\partial\beta} = -\frac{\partial}{\partial\beta}\left(\left.\frac{d\log g(u)}{du}\right|_{u=y_j-x_j^\top\beta}\right)x_j^\top = x_j x_j^\top \cdot \left.\frac{d^2\log g(u)}{du^2}\right|_{u=y_j-x_j^\top\beta}.$$

Hence we have

$$-\frac{\partial^2\ell}{\partial\beta^2} = -\sum_j x_j x_j^\top \left.\frac{d^2\log g(u)}{du^2}\right|_{u=y_j-x_j^\top\beta}$$

leading to the Fisher information matrix

$$I(\beta) = \mathbb{E}\left\{-\frac{\partial^2\ell}{\partial\beta^2}\right\} = \sum_{j=1}^{n} x_j x_j^\top \mathbb{E}\left\{-\left.\frac{d^2\log g(u)}{du^2}\right|_{u=y_j-x_j^\top\beta}\right\},$$

where the expectation is found to be exactly

$$\int -\frac{d^2\log g(u)}{du^2} g(u)\,du = i_g.$$

Hence we have $I(\beta) = i_g X^\top X$ and so the asymptotic variance of the MLE estimator is $i_g^{-1}(X^\top X)^{-1}$. The asymptotic relative least squares efficiency with respect to the maximum likelihood estimate is therefore

$$\left\{\frac{|i_g^{-1}(X^\top X)^{-1}|}{|\sigma^2(X^\top X)^{-1}|}\right\}^{1/p} = \frac{1}{i_g\sigma^2}.$$

(a) With $g(u) = (2\pi\sigma^2)^{-1/2}e^{-u^2/(2\sigma^2)}$ for $u \in \mathbb{R}$, it is $i_g = 1/\sigma^2$, so the efficiency is 1. This is not surprising, since the least squares estimator and MLE estimator coincide in the case of Gaussian distribution.

(c) Let $\lambda = \sqrt{2}/\sigma$, where σ^2 is the variance. The density of the Laplace distribution can be written with this parametrization as $g(u) = \frac{\lambda}{2}\exp(-\lambda|u|)$ for $u \in \mathbb{R}$. Since the regularity conditions for MLE are satisfied in this case (no need to verify it), we have

$$i_g = \int -\frac{d^2\log g(u)}{du^2} g(u)\,du = \int \left\{\frac{d\log g(u)}{du}\right\}^2 g(u)\,du = \int \{-\lambda\,\mathrm{sgn}(u)\}^2 (\lambda/2)\exp(-\lambda|u|)\,du = \lambda^2.$$

Thus the relative asymptotic efficiency is $1/(\lambda^2\sigma^2) = 1/2$.

Problem 44. Give the equivalent of the H matrix for non-parametric regression with kernel smoothing.

Solution
The adjusted values are

$$\widehat{y}_i = \widehat{g}(x_i) = \sum_{j=1}^{n} y_j \frac{K\left(\frac{x_j - x_i}{\lambda}\right)}{\sum_{k=1}^{n} K\left(\frac{x_k - x_i}{\lambda}\right)}.$$

By defining

$$S_{\lambda,ij} = \frac{K\left(\frac{x_j - x_i}{\lambda}\right)}{\sum_{k=1}^{n} K\left(\frac{x_k - x_i}{\lambda}\right)},$$

we have $\widehat{y} = S_\lambda y$, where S_λ is called the *smoother matrix*. In non-parametric regression, this is the analogue of the hat matrix.

Problem 45. (Cubic spline)
Let $n \geq 2$ and $a < x_1 < x_2 < \cdots < x_n < b$. Denote by $N(x_1, x_2, \ldots, x_n)$ the space of natural cubic splines with knots $x_1, x_2, \ldots, x_n$. The goal of this exercise is to show that the solution to the problem

$$\min_{f \in C^2[a,b]} L(f), \text{ où } L(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b \{f''(x)\}^2 dx, \quad \lambda > 0, \tag{14}$$

must belong to $N(x_1, x_2, \ldots, x_n)$. In order to show this, we need the following theorem

> **Theorem.** For every set of points $(x_1, z_1), (x_2, z_2), \ldots, (x_n, z_n)$, it exists a natural cubic spline g interpolating those points. In other words, $g(x_i) = z_i$, $i = 1, \ldots, n$, for a unique natural cubic spline g. Moreover, the knots of g are $x_1, x_2, \ldots, x_n$.

1. Let g the natural cubic spline interpolating the points (x_i, z_i), $i = 1, \ldots, n$, and let $\widetilde{g} \in C^2[a,b]$ another function interpolating the same points. Show that

$$\int_a^b g''(x) h''(x) dx = 0,$$

   where $h = \widetilde{g} - g$.
   Hint: integration by parts

2. Using point (1) show that

$$\int_a^b \{\widetilde{g}''(x)\}^2 dx \geq \int_a^b \{g''(x)\}^2 dx$$

   when the equality holds if and only if $\widetilde{g} = g$.

3. Use point (2) to show that if the problem (14) has a solution $\widehat{f}$, then $\widehat{f} \in N(x_1, x_2, \ldots, x_n)$.

Solution

1. Using integration by parts, we obtain that

$$\begin{aligned}
\int_a^b g''(x) h''(x) dx &= \underbrace{g''(x) h'(x)\Big|_a^b}_{=0, \text{ car } g''(a)=g''(b)=0} - \int_a^b g'''(x) h'(x) dx \\
&= -\sum_{i=1}^{n-1} g'''(x_i^+) \int_{x_i}^{x_{i+1}} h'(x) dx \\
&= -\sum_{i=1}^{n-1} g'''(x_i^+) \{h(x_{i+1}) - h(x_i)\} = 0.
\end{aligned}$$

Here, the second equality comes from the fact that $g'''(x) = 0$ inside the intervals (a, x_1) and (x_n, b) and that $g'''(x)$ equals to the constant $\lim_{x\to x_i^+} g'''(x) = g'''(x_i^+)$ inside the interval (x_i, x_{i+1}). To obtain the last equality finally, observe that $\widetilde{g}(x_i) = g(x_i) = z_i$ hence $h(x_i) = 0$ for every i.

2. By direct computation we obtain that

$$\begin{aligned}\int_a^b \{\widetilde{g}''(x)\}^2 dx &= \int_a^b \{g''(x) + h''(x)\}^2 dx \\ &= \int_a^b \{g''(x)\}^2 dx + 2\int_a^b g''(x)h''(x)dx + \int_a^b \{h''(x)\}^2 dx \\ &= \int_a^b \{g''(x)\}^2 dx + \int_a^b \{h''(x)\}^2 dx \geq \int_a^b \{g''(x)\}^2 dx.\end{aligned}$$

where we have equality if and only if $h''(x) \equiv 0$, so we must have $h(x) = kx + c$. But since $h(x_i) = 0$ for every i, it must be that$h(x) \equiv 0$. In particular we have equality if and only if $\widetilde{g} = g$.

3. Let $\widetilde{f} \in C^2[a,b] \backslash N(x_1, \ldots, x_n)$ and let $f \in N(x_1, \ldots, x_n)$ the spline which is interpolating the points $(x_i, \widetilde{f}(x_i))$, $i = 1, \ldots, n$. The existance of f is guaranteed by the theorems seen in class. By point (2)

$$\int_a^b \{\widetilde{f}''(x)\}^2 dx > \int_a^b \{f''(x)\}^2.$$

Moreover

$$\sum_{i=1}^n (y_i - \widetilde{f}(x_i))^2 = \sum_{i=1}^n (y_i - f(x_i))^2.$$

Hence, $L(\widetilde{f}) > L(f)$ and we notice that if the minimum exists, it must belong to $N(x_1, \ldots, x_n)$.

Remark. Using the properties of splines, it it possible to show that a minimum always exists and is unique. Hence the problem $\min_{f\in C^2[a,b]} L(f)$ admits always a unique solution and this solution is a natural cubic spline.

Compiled 09/12/2019 at 18:51