

Midterm 1: Math 6266 (Zhilova)

Peter Williams

Section 1.1

Exercise 1.

Consider the linear regression model with mean zero, uncorrelated, heteroscedastic noise:

$$Y_i = X_i^\top \theta + \varepsilon_i, \text{ for } i = 1, \dots, n, \quad E\varepsilon_i = 0, \quad \text{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma_i^2, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \quad (1)$$

Find expressions for the LSE and response estimator in this model

To set up the problem, take $W^{-1} = \text{diag}\{\sigma_1^2, \dots, \sigma_n^2\}$, $W = \text{diag}\{\frac{1}{\sigma_1^2}, \dots, \frac{1}{\sigma_n^2}\}$, $W^{1/2} = \text{diag}\{\sqrt{\frac{1}{\sigma_1^2}}, \dots, \sqrt{\frac{1}{\sigma_n^2}}\}$, with $W^\top = W$, and $W^{1/2}W^{1/2} = W$, since they are diagonal matrices. Also we will use $w_i = \frac{1}{\sigma_i^2} = W_{ii}$.

Under heteroscedastic noise assumptions, we define the least squares estimator, denoted $\hat{\theta}$, as:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^n w_i (Y_i - X_i^\top \theta)^2 = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^n (\sqrt{w_i} Y_i - \sqrt{w_i} X_i^\top \theta)^2 = \underset{\theta}{\operatorname{argmin}} \|W^{1/2} Y - W^{1/2} X^\top \theta\|^2$$

$$G(\theta) = \|W^{1/2} Y - W^{1/2} X^\top \theta\|^2 = (W^{1/2} Y - W^{1/2} X^\top \theta)^\top (W^{1/2} Y - W^{1/2} X^\top \theta) = Y^\top W Y - 2\theta^\top X W Y + \theta^\top X W X^\top \theta$$

with gradient,

$$\nabla G(\theta) = -2XWY + 2XWX^\top \theta$$

Setting this expression equal to zero leads to estimator $\hat{\theta} = (XWX^\top)^{-1}XWY$, which leads to response estimator $\hat{Y} = X^\top \hat{\theta} = X^\top (XWX^\top)^{-1}XWY$.

Exercise 2.

Assume that $\varepsilon_i \sim N(0, \sigma_i^2)$ in the previous problem. What is known about the distribution of $\hat{\theta}$ and \hat{Y} ?

For $\hat{\theta}$, we have,

$$E[\hat{\theta}] = E[(XWX^\top)^{-1}XWY] = E[(XWX^\top)^{-1}XW(X^\top \theta^* + \varepsilon)] = E[\theta^*] + E[(XWX^\top)^{-1}XW\varepsilon] = \theta^*$$

indicating that $\hat{\theta}$ is unbiased. Further $\hat{\theta}$ is normally distributed, since is a linear transformation of $\varepsilon \sim N(0, W^{-1})$. Further we have,

$$\begin{aligned} \text{Var}(\hat{\theta}) &= \text{Var}((XWX^\top)^{-1}XWY) = \text{Var}((XWX^\top)^{-1}XW(X^\top \theta^* + \varepsilon)) = \text{Var}((XWX^\top)^{-1}XW\varepsilon) = \dots \\ &= (XWX^\top)^{-1}XW \text{Var}(\varepsilon) W^\top X^\top (XWX^\top)^{-1} = (XWX^\top)^{-1}XWX^\top (XWX^\top)^{-1} = (XWX^\top)^{-1} = \text{Var}(\hat{\theta}) \end{aligned}$$

For \hat{Y} we have,

$$E[\hat{Y}] = E[X^\top (XWX^\top)^{-1}XWY] = E[X^\top (XWX^\top)^{-1}XW(X^\top \theta^* + \varepsilon)] = E[X^\top \theta^* + X^\top (XWX^\top)^{-1}XW\varepsilon] = E[X^\top \theta^*] = Y$$

and,

$$\begin{aligned} \text{Var}[\hat{Y}] &= \text{Var}[X^\top (XWX^\top)^{-1}XWY] = \text{Var}[X^\top (XWX^\top)^{-1}XW(X^\top \theta^* + \varepsilon)] = \text{Var}[X^\top \theta^* + X^\top (XWX^\top)^{-1}XW\varepsilon] = \dots \\ &= \text{Var}[X^\top (XWX^\top)^{-1}XW\varepsilon] = X^\top (XWX^\top)^{-1}XW \text{Var}(\varepsilon) W^\top X^\top (XWX^\top)^{-1}X = \dots \end{aligned}$$

$$= X^\top(XWX^\top)^{-1}XWX^\top(XWX^\top)^{-1}X = X^\top(XWX^\top)^{-1}X = \text{Var}[\hat{Y}]$$

Now suppose additionally that $\sigma_i^2 \equiv \sigma^2 > 0$. What can be said about distribution of the estimator $\hat{\sigma}^2$?

With $\sigma_i^2 \equiv \sigma^2 > 0$, we have $\hat{\sigma}^2 = \frac{\|Y - X^\top \hat{\theta}\|^2}{n-p} = \frac{\|\hat{\varepsilon}\|^2}{n-p}$. Further denote, $\|\hat{\varepsilon}\| = \|Y - \hat{Y}\| = \|Y - \Pi Y\| = \|(I_n - \Pi)Y\|$, also noting that $(I_n - \Pi)X^\top = X^\top - \Pi X^\top = X^\top - X^\top(XX^\top)^{-1}XX^\top = X^\top - X^\top = 0$.

Then we have,

$$\begin{aligned} (n-p)E[\hat{\sigma}^2] &= E[\|Y - X^\top \hat{\theta}\|^2] = E[\|\hat{\varepsilon}\|^2] = E[\text{tr}(\hat{\varepsilon}\hat{\varepsilon}^\top)] = E[\text{tr}((I_n - \Pi)YY^\top(I_n - \Pi))] = \dots \\ &= E[\text{tr}((I_n - \Pi)(X^\top \theta^* + \varepsilon)(X^\top \theta^* + \varepsilon)^\top(I_n - \Pi))] = E[\text{tr}((I_n - \Pi)\varepsilon\varepsilon^\top(I_n - \Pi))] = \text{tr}((I_n - \Pi)E[\varepsilon\varepsilon^\top]) = \dots \end{aligned}$$

Using the cyclic property of the trace operator, the property that $(I_n - \Pi)(I_n - \Pi) = (I_n - \Pi)$, and the expectation $E[\varepsilon\varepsilon^\top] = \sigma^2 I_n$, leading to

$$\dots = \sigma^2 \text{tr}(I_n - \Pi) = \sigma^2(n-p) = (n-p)E[\hat{\sigma}^2]$$

Looking further at the distribution of $\|Y - X^\top \hat{\theta}\|^2 = \hat{\varepsilon}^\top \hat{\varepsilon}$, we have

$$\hat{\varepsilon}^\top \hat{\varepsilon} = ((I_n - \Pi)Y)^\top((I_n - \Pi)Y) = Y^\top(I_n - \Pi)Y = (X^\top \theta^* + \varepsilon)^\top(I_n - \Pi)(X^\top \theta^* + \varepsilon) = \varepsilon^\top(I_n - \Pi)\varepsilon$$

Since we know that $\varepsilon \sim N(0, \sigma^2 I_n)$, and further $\frac{\varepsilon^\top \varepsilon}{\sigma^2} \sim \chi^2(n)$, $(\frac{\varepsilon}{\sigma})^\top(I_n - \Pi)(\frac{\varepsilon}{\sigma}) \sim \chi^2(n-p)$, since we know from earlier that $(I_n - \Pi)$, is idempotent, with rank equal to $\text{tr}(I_n - \Pi) = \text{tr}(I_n) - \text{tr}(\Pi) = n - p$.

Exercise 3.

Consider the linear regression model from exercise 1. Suppose, that the target of estimation is $\gamma^\top \theta$ for some determinate non-zero vector $\gamma \in R^p$. Find expression for the LSE of $\gamma^\top \theta$. Is this estimate optimal in sense of Gauss-Markov theorem, i.e. does it have the smallest variance among all linear unbiased estimators?

Using our findings from exercise 2, we have an unbiased LSE estimator in $\gamma^\top \hat{\theta}$ since $E[\gamma^\top \hat{\theta} - \gamma^\top \theta] = \gamma^\top E[(XWX^\top)^{-1}XWY] - \gamma^\top \theta = \gamma^\top \theta - \gamma^\top \theta = 0$.

Using another finding from exercise 2 we have, $\text{Var}(\gamma^\top \hat{\theta}) = \gamma^\top \text{Var}(\hat{\theta})\gamma = \gamma^\top (XWX^\top)^{-1}\gamma$.

To show that $\gamma^\top \hat{\theta}$ is BLUE, we compare it to, to another estimator $\tilde{\theta} = ((XWX^\top)^{-1}XW + D)Y$, where D is a $p \times n$ matrix. We then have

$$\begin{aligned} E[\tilde{\theta}] &= E[((XWX^\top)^{-1}XW + D)Y] = E[((XWX^\top)^{-1}XW + D)(X^\top \theta + \varepsilon)] = \dots \\ &= E[\theta] + E[DX^\top \theta] + E[(XWX^\top)^{-1}XW\varepsilon] = \theta + DX^\top \theta + 0 \end{aligned}$$

So $\tilde{\theta}$ is only an unbiased estimator when $DX^\top = 0$.

The variance of $\tilde{\theta}$ is:

$$\begin{aligned} \text{Var}(\tilde{\theta}) &= \text{Var}[((XWX^\top)^{-1}XW + D)Y] = \\ &= [((XWX^\top)^{-1}XW + D)]\text{Var}(Y)[((XWX^\top)^{-1}XW + D)^\top] = \dots \\ &= [((XWX^\top)^{-1}XW + D)]\text{Var}(X^\top \theta^* + \varepsilon)[(WX^\top(XWX^\top)^{-1} + D^\top)] = \dots \\ &= [((XWX^\top)^{-1}XW + D)]W^{-1}[(WX^\top(XWX^\top)^{-1} + D^\top)] = \dots \\ &= [((XWX^\top)^{-1}XW + D)][(X^\top(XWX^\top)^{-1} + W^{-1}D^\top)] = \dots \\ &= (XWX^\top)^{-1} + DW^{-1}D^\top + DX^\top(XWX^\top)^{-1} + (XWX^\top)^{-1}XD^\top = \text{Var}(\hat{\theta}) \end{aligned}$$

But in order for our estimator to be unbiased, we have $DX^\top = XD^\top = 0$. Therefore we have:

$$\text{Var}(\tilde{\theta}) = (XWX^\top)^{-1} + DW^{-1}D^\top$$

Finally, taking $\gamma \in R^p$ we have $\text{Var}(\gamma^\top \tilde{\theta}) = \gamma^\top \text{Var}(\tilde{\theta}) \gamma = \gamma^\top ((XWX^\top)^{-1} + DW^{-1}D^\top) \gamma$. In comparison, our LSE estimator $\hat{\theta}$, has $\text{Var}(\gamma^\top \hat{\theta}) = \gamma^\top ((XWX^\top)^{-1}) \gamma$.

We have then

$$\text{Var}(\gamma^\top \tilde{\theta}) = \gamma^\top \text{Var}(\tilde{\theta}) \gamma = \gamma^\top ((XWX^\top)^{-1} + DW^{-1}D^\top) \gamma \geq \text{Var}(\gamma^\top \hat{\theta}) = \gamma^\top ((XWX^\top)^{-1}) \gamma$$

Since W^{-1} is a diagonal matrix with all positive elements, DD^\top is symmetric positive, the form $\gamma^\top DW^{-1}D^\top \gamma \geq 0$, implying $\text{Var}(\gamma^\top \tilde{\theta}) \geq \text{Var}(\gamma^\top \hat{\theta})$.

Section 1.3

Exercise 4.

Let $A \in R^{n \times n}$ be a matrix (corresponding to a linear map in R^n). Show that A preserves length for all $x \in R^n$ iff it preserves the inner product. I.e. one needs to show the following:

$$\|Ax\| = \|x\| \quad \forall x \in R^n \iff (Ax)^\top (Ay) = x^\top y \quad \forall x, y \in R^n.$$

Take,

$$\|x\| = \sqrt{x \cdot x} = \sqrt{x^\top x} \implies \|Ax\| = \sqrt{Ax \cdot Ax} = \sqrt{x^\top A^\top A x} \implies$$

,

$$A^\top A = I_n = A^{-1}, \quad A^\top = A^{-1}, \quad \|Ax\| = \|x\|$$

this implies A is an orthogonal matrix, and further,

$$(Ax)^\top (Ay) = \|Ax Ay\|^2 = x^\top A^\top A y = x^\top y = \|xy\|^2$$

Exercise 5.

(a) Let $x_0 \in R^n$ be some fixed vector, find a projection map on the subspace $\text{span}(x_0)$. Compare your result with matrix Π (from section 1.3) for the case of $p = 1$.

Let $x = \text{span}(x_0) = \text{span}(x_1, x_2, \dots, x_n)$, denote the subspace of interest, and x_1, x_2, \dots are basis vectors and $y = (y_1, y_2, \dots, y_n)^\top$. The projection map is,

$$\text{Proj}_x(y) = \frac{y \cdot x}{y \cdot y} x = \sum_{i=1}^n \frac{y_i \cdot x_i}{y_i \cdot y_i} x_i$$

For the case $p = 1$, and $\Pi = X(X^\top X)^{-1}X^\top$, $X^\top \in R^n$, we have,

$$\Pi y = \hat{y} = X^\top (X X^\top)^{-1} X y = X^\top \frac{X y}{X X^\top} = \frac{\sum_i x_i y_i}{\sum_i x_i^2} (x_1, x_2, \dots, x_n)^\top = \frac{\langle X \cdot y \rangle}{\langle y \cdot y \rangle} X^\top = \text{Proj}_X(y)$$

(b) Prove part 3) of Lemma 1.1 for an arbitrary orthogonal projection in R^n . Show $\forall h \in R^n$, $\|h\|^2 = \|\Pi h\|^2 + \|h - \Pi h\|^2$.

Using the fact that $(I_n - \Pi)^\top (I_n - \Pi) = I_n - 2\Pi + \Pi = I_n - \Pi$, we have,

$$\|h\|^2 = \|\Pi h\|^2 + \|h - \Pi h\|^2 = h^\top \Pi^\top \Pi h + h^\top (I_n - \Pi)^\top (I_n - \Pi) h = h^\top \Pi h + h^\top (I_n - \Pi) h = h^\top I_n h + h^\top \Pi h - h^\top \Pi h = \|h\|^2$$

Exercise 6.

Let L_1, L_2 be some subspaces in R^n , and $L_2 \subseteq L_1 \subseteq R^n$. Let P_{L_1}, P_{L_2} denote orthogonal projections on these subspaces. Prove the following properties:

(a) $P_{L_2} - P_{L_1}$ is an orthogonal projection,

Denote L_1 as a subset of R^n with orthonormal basis $\text{span}\{u_1, u_2, \dots, u_p\}$, and L_2 with basis $\text{span}\{u_1, u_2, \dots, u_{p-k}\} \subseteq \text{span}\{u_1, \dots, u_p\}$. For a vector $x \in R^n$, we have an orthogonal projection onto L_1 and L_2 denoted as follows:

$$P_{L_1}(x) = \sum_{i=1}^p (x \cdot u_i) u_i, \quad P_{L_2}(x) = \sum_{i=1}^{p-k} (x \cdot u_i) u_i$$

The difference of these projections is then:

$$P_{L_2}(x) - P_{L_1}(x) = (P_{L_2} - P_{L_1})x = \sum_{i=1}^{p-k} (x \cdot u_i) u_i - \sum_{i=1}^p (x \cdot u_i) u_i = (-1) \cdot \sum_{i=p-k+1}^p (x \cdot u_i) u_i$$

which is an orthogonal projection onto the subspace, defined as $\text{span}\{u_{p-k+1}, u_{p-k+2}, \dots, u_p\} \subseteq \text{span}\{u_1, \dots, u_p\}$.

(b) $\|P_{L_2}x\| \leq \|P_{L_1}x\| \quad \forall x \in R^n$,

We have $\|P_{L_2}x\| = \|\sum_{i=1}^{p-k} (x \cdot u_i) u_i\|$ and $\|P_{L_1}x\| = \|\sum_{i=1}^p (x \cdot u_i) u_i\|$. For $k < p$, we have

$$\|P_{L_1}(x) - P_{L_2}(x)\| = \left\| \sum_{i=p-k+1}^p (x \cdot u_i) u_i \right\| \geq 0,$$

and by the triangle inequality,

$$\|P_{L_2}x\| \leq \|P_{L_1}(x)\| = \|(P_{L_1}x - P_{L_2}x) + P_{L_2}x\| \leq \|P_{L_1}x - P_{L_2}x\| + \|P_{L_2}x\|$$

(c) $P_{L_2} \cdot P_{L_1} = P_{L_2}$

We can denote $P_{L_1}(x) = \sum_{i=1}^p (x \cdot u_i) u_i = UU^T x$, where matrix $U_{n \times p}$ consists of orthonormal vectors $[u_1, \dots, u_p]$, and denote

$$P_{L_2}(x) = \sum_{i=1}^{p-k} (x \cdot u_i) u_i = VV^T x$$

where matrix $V_{n \times (p-k)}$ consists of orthonormal vectors $[u_1, \dots, u_{p-k}]$. So the product $P_{L_2}P_{L_1}$ can be written

$$P_{L_2}P_{L_1} = VV^TUU^T$$

Since the first $p-k$ column vectors of V and U are the same, and orthonormal, the inner product $V^T U$ generates a $(p-k) \times p$ block matrix of the form $\begin{bmatrix} I_{p-k} & 0 \end{bmatrix}$ where 0 is a $k \times k$ matrix of zeroes. We then have

$$P_{L_2}P_{L_1} = VV^TUU^T = V \begin{bmatrix} I_{p-k} & 0 \end{bmatrix} U^T = VV^T = P_{L_2}$$

Section 2.1

Exercise 8.

Let $X \sim N(0, I_n)$, $Q = X^T X$. Suppose that Q is decomposed into the sum of two quadratic forms: $Q = Q_1 + Q_2$, where $Q_i = X^T A_i X$, $i = 1, 2$ for some symmetric matrices A_1, A_2 with $\text{rank}(A_1) = n_1$ and $\text{rank}(A_2) = n_2$. Show that if $n_1 + n_2 = n$, then Q_1 and Q_2 are independent and $Q_i \sim \chi^2(n_i)$ for $i = 1, 2$.

First note that $X^\top X \sim \chi^2(n)$, since $X^\top X = \sum_{i=1}^n x_i^2$, which is the sum of iid squared normal random variables with variance 1.

Since A_1 is a symmetric matrix, we can diagonalize it, $A_1 = U^\top \Lambda U$. We know the rank of A_1 is n_1 . This implies that $U^\top A_1 U = \Lambda = \text{diag}\{\Lambda_1, \dots, \Lambda_{n_1}, \dots, \Lambda_n\}$, has n_1 non-zero, positive eigenvalues, and n_2 eigenvalues that equal zero.

Using the orthogonal matrix U from the decomposition of A_1 , we set $X = UY$, so that $X^\top X = Y^\top U^\top U Y = Y^\top I_n Y = Y^\top Y$. So $Q = X^\top X = Y^\top Y = \sum_{i=1}^n Y_i^2$.

We can write

$$Q = Q_1 + Q_2 = \sum_{i=1}^n Y_i^2 = Y^\top U^\top A_1 U Y + Y^\top U^\top A_2 U Y = Y^\top \Lambda Y + Y^\top U^\top A_2 U Y = \sum_{i=1}^n \Lambda_i Y_i^2 + Y^\top U^\top A_2 U Y$$

Since only n_1 eigenvalues in Λ are non-zero, we have

$$Q = \sum_{i=1}^{n_1} \Lambda_i Y_i^2 + \sum_{i=n_1+1}^n \Lambda_i Y_i^2 + Y^\top U^\top A_2 U Y = Q = \sum_{i=1}^{n_1} \Lambda_i Y_i^2 + Y^\top U^\top A_2 U Y$$

,

if we organize Λ in way such that the positive eigenvalues on the diagonal are present in the first n_1 diagonal elements. So we have $Q_1 = \sum_{i=1}^{n_1} \Lambda_i Y_i^2$

To solve for $Q_2 = X^\top X = Y^\top U^\top A_2 U Y$, from above we have

$$Y^\top U^\top A_2 U Y = Q - Q_1 = Q - \sum_{i=1}^{n_1} \Lambda_i Y_i^2 = \sum_{i=1}^{n_1} Y_i^2 + \sum_{i=n_1+1}^n Y_i^2 - \sum_{i=1}^{n_1} \Lambda_i Y_i^2 = \sum_{i=1}^{n_1} (1 - \Lambda_i) Y_i^2 + \sum_{i=n_1+1}^n Y_i^2$$

We know the rank of A_2 is $n_2 = n - n_1$. So the term $\sum_{i=1}^{n_1} (1 - \Lambda_i) Y_i^2$ must equal zero, implying that $\Lambda_1 = \Lambda_2 = \dots = \Lambda_{n_1} = 1$. This also implies $Q = Q_1 + Q_2 = \sum_{i=1}^{n_1} Y_i^2 + \sum_{i=n_1+1}^n Y_i^2$.

Since each squared element $Y_i^2 = X_i^2 \sim \chi^2(1)$ in Q only occurs once in the summand, we can say that and $Q_1 = \sum_{i=1}^{n_1} Y_i^2 \sim \chi^2(n_1)$, and $Q_2 = \sum_{i=n_1+1}^n Y_i^2 \sim \chi^2(n_2)$, since $Q = Q_1 + Q_2 \sim \chi^2(n)$.

Section 2.2

Exercise 9.

In the Gaussian linear regression model 3, consider the target of estimation $\eta = H^\top \theta^*$, where $H \in R^{q \times p}$ is some non-zero matrix with $q \leq p$. Find an analogue of the quadratic form S_2 (from (4)) for the new target η^* , and prove for the new quadratic form statements similar to (e) from Theorem 2.1, and Corollary 2.1.2.

With $\eta^* = H^\top \theta^*$, and $\hat{\eta} = H^\top \hat{\theta}$, we have,

$$E[\hat{\eta}] = E[H^\top \hat{\theta}] = H^\top E[\hat{\theta}] = H^\top E[(X X^\top)^{-1} X Y] = H^\top E[(X X^\top)^{-1} X (X^\top \theta^* + \varepsilon)] = H^\top \theta^*$$

and

$$\begin{aligned} \text{Var}(H^\top \hat{\theta}) &= H^\top \text{Var}(\hat{\theta}) H = H^\top \text{Var}((X X^\top)^{-1} X (X^\top \theta^* + \varepsilon)) H = H^\top \text{Var}(\theta^* + (X X^\top)^{-1} X \varepsilon) H = \dots \\ &= H^\top ((X X^\top)^{-1} X \sigma^2 I_n X^\top (X X^\top)^{-1} H = \sigma^2 H^\top (X X^\top)^{-1} H = \sigma^2 S = \text{Var}(H^\top \hat{\theta}) \end{aligned}$$

Since $H^\top \hat{\theta}$ is a linear transformation of normal random variables, we have,

$$\frac{H^\top \hat{\theta} - H^\top \theta^*}{\sqrt{\sigma^2 H^\top (X X^\top)^{-1} H}} = \frac{\hat{\eta} - \eta^*}{\sigma \sqrt{S}} \sim N(0, I_p)$$

We can then have an analog of S_2 from theorem 2.1:

$$\frac{\|S^{-1/2}(H^\top \hat{\theta} - H^\top \theta^*)\|^2}{\sigma^2} = \frac{\|S^{-1/2}(\hat{\eta} - \eta^*)\|^2}{\sigma^2} = \frac{(\hat{\eta} - \eta^*)^\top (S^{-1})(\hat{\eta} - \eta^*)}{\sigma^2} \sim \chi^2(p)$$

Exercise 10.

(a) Consider model (3) for $p = 2$, $X_i = (1, x_i)^\top$, $\theta^* = (\theta_1^*, \theta_2^*)^\top$ (similarly to section 1.5). Write explicit expressions for the confidence sets for θ^* , θ_1^* , θ_2^* .

To set up explicit expression for the case above, we have:

$$XX^\top = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

and $\det(XX^\top) = n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 = n \sum_{i=1}^n (x_i - \bar{x})^2$, and

$$(XX^\top)^{-1} = \frac{n}{\det(XX^\top)} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

So we have

$$\begin{aligned} \hat{\theta} &= (XX^\top)^{-1}XY = \frac{n}{\det(XX^\top)} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} = (\hat{\theta}_1, \hat{\theta}_2)^\top = \dots \\ &\dots = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \bar{y} \sum_i x_i^2 - \bar{x} \sum_i x_i y_i \\ \sum_i x_i y_i - n \bar{y} \bar{x} \end{bmatrix} = (\hat{\theta}_1, \hat{\theta}_2)^\top = \hat{\theta} \end{aligned}$$

To find a confidence region for θ^* , using a mixture of matrix and summation notation, we use the property:

$$\frac{\|(XX^\top)^{1/2}(\hat{\theta} - \theta^*)\|^2}{\sum_{i=1}^n (y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i)^2} \frac{n-2}{2} \sim F(2, n-2)$$

and denote $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i)^2}{n-2}$. Where F denotes the F distribution with $df_1 = 2$, and $df_2 = n-2$.

We can create a confidence interval for θ^* , such that, qF_α denotes the α^{th} quantile for $F(2, n-2)$.

$$P\left(\frac{\|(XX^\top)^{1/2}(\hat{\theta} - \theta^*)\|^2}{p\hat{\sigma}^2} < qF_{1-\alpha}\right) = 1 - \alpha = P((\hat{\theta} - \theta^*)^\top \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} (\hat{\theta} - \theta^*) < p\hat{\sigma}^2 qF_{1-\alpha})$$

We know that $\frac{(XX^\top)^{1/2}(\hat{\theta} - \theta^*)}{\hat{\sigma}} \sim N(0, I_p)$. We can then set up confidence intervals for θ_1^* and θ_2^* .

For θ_1^* , we can set up a T -statistic by taking the difference of the first parameter estimate and the true estimate and dividing it the corresponding standard error:

$$T_{1(n-2-1)} = \frac{\hat{\theta}_1 - \theta_1^*}{\sqrt{\hat{\sigma}^2 [(XX^\top)^{-1}]_{11}}} = \frac{\hat{\theta}_1 - \theta_1^*}{\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i)^2}{n-p} \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

Using T_1 we can set up a % $100(1 - \alpha)$ confidence interval for $\hat{\theta}_1^*$ via:

$$\hat{\theta}_1^* \pm T_{1(n-3), \alpha/2} \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i)^2}{n-p} \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

For θ_2^* we have:

$$T_{2(n-3)} = \frac{\hat{\theta}_2 - \theta_2^*}{\sqrt{\hat{\sigma}^2 [(XX^\top)^{-1}]_{22}}} = \frac{\hat{\theta}_2 - \theta_2^*}{\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i)^2}{n-p} \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

With T_2 we can set up a % 100(1 - α) confidence interval for $\hat{\theta}_2^*$ via:

$$\theta_2^* \pm T_{2(n-3), \alpha/2} \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i)^2}{(n-p) \sum_{i=1}^n (x_i - \bar{x})^2}}$$

(b) Find a confidence interval for the expected response $E[Y_i]$ in the model in part (a). The variance of the expected response $\text{var}(\hat{Y}) = \text{var}(X^\top (XX^\top)^{-1} XY) = \text{var}(X^\top (XX^\top)^{-1} X(X^\top \theta^* + \varepsilon)) = \text{var}(X^\top (XX^\top)^{-1} X \varepsilon) = \sigma^2 X^\top (XX^\top)^{-1} X$. Using the standard error for \hat{Y} , we can set up up the following confidence interval for the expected response for the i^{th} record using a T-statistic:

$$T_{(n-3)} = \frac{\hat{y}_i - y_i}{\sqrt{\hat{\sigma}^2 x_i^\top (XX^\top)^{-1} x_i}} = \frac{\hat{y}_i - y_i}{\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i)^2}{n-2} x_i^\top (XX^\top)^{-1} x_i}}$$

With this statistic a % 100(1 - α) confidence interval for y_i can be created:

$$y_i \pm T_{n-3, \alpha/2} \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i)^2}{n-2} x_i^\top \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} x_i}$$

Exercise 11.

Find an elliptical confidence set for the expected response $E[Y]$ in model (3).

For the model $Y = X^\top \theta^* + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I_n)$, with $\hat{Y} = X^\top \hat{\theta} = X^\top (XX^\top)^{-1} XY = \Pi Y$, we have

$$E(\hat{Y} - Y) = E(\hat{Y}) - E[Y] = E[X^\top (XX^\top)^{-1} X(X^\top \theta^* + \varepsilon)] - X^\top \theta^* - \varepsilon = E[X^\top \theta^*] - E[X^\top \theta^*] - E[\varepsilon] = 0$$

and

$$\text{Var}(\hat{Y} - Y) = \text{Var}((\Pi - I_n)Y) = (\Pi - I_n) \text{Var}(X^\top \theta^* + \varepsilon) (\Pi - I_n)^\top = (\Pi - I_n) \sigma^2 I_n (\Pi - I_n)^\top = \sigma^2 (I_n - \Pi)$$

since \hat{Y} is a linear transformation of a gaussian, we have $\hat{Y} \sim N(0, \sigma^2 (I_n - \Pi))$. Setting $S = \sigma^2 (I_n - \Pi) = \text{Var}(\hat{Y} - Y)$, we have,

$$\frac{(\hat{Y} - Y) - 0}{\sqrt{\sigma^2 (I_n - \Pi)}} = \frac{(\hat{Y} - Y) - 0}{S^{1/2}} \sim N(0, I_{n-p})$$

This implies,

$$\|S^{-1/2}(\hat{Y} - Y)\|^2 = (\hat{Y} - Y) S^{-1} (\hat{Y} - Y)^\top \sim \chi^2(n-p)$$

Using this information we can set up confidence region (known σ^2), non-pivotal, for \hat{Y} with an elliptical region satisfying for some constant c , and significance level α , in the region around Y ,

$$P((\hat{Y} - Y) S^{-1} (\hat{Y} - Y)^\top \leq c^2) = 1 - \alpha$$