Midterm 1: Math 6266 (Zhilova)

Peter Williams

Section 1.1

Exercise 1.

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

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

Find expressions for the LSE and response estimator in this model

To set up the problem, take $W^{-1} = diag\{\sigma_1^2, ..., \sigma_n^2\}$, $W = diag\{\frac{1}{\sigma_1^2}, ..., \frac{1}{\sigma_n^2}\}, W^{1/2} = diag\{\sqrt{\frac{1}{\sigma_1^2}}, ..., \sqrt{\frac{1}{\sigma_n^2}}\}$, with $W^\intercal = W$, 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:

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

$G(\theta) = ||W^{1/2}Y - W^{1/2}X^\intercal\theta||^2 = (W^{1/2}Y - W^{1/2}X^\intercal\theta)^\intercal(W^{1/2}Y - W^{1/2}X^\intercal\theta) = Y^\intercal WY - 2\theta^\intercal XWY + \theta^\intercal XWX^\intercal\theta$

with gradient,

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

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

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,

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

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,

$$Var(\hat{\theta}) = Var((XWX^\intercal)^{-1}XWY) = Var((XWX^\intercal)^{-1}XW(X^\intercal\theta^* + \varepsilon)) = Var((XWX^\intercal)^{-1}XW\varepsilon)) = ...$$

$$= (XWX^\intercal)^{-1}XWVar(\varepsilon)W^\intercal X^\intercal(XWX^\intercal)^{-1} = (XWX^\intercal)^{-1}XWX^\intercal(XWX^\intercal)^{-1} = (XWX^\intercal)^{-1} = Var(\hat{\theta})$$

For $\hat{Y}$ we have,

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

and,

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

$$... = Var[X^\intercal(XWX^\intercal)^{-1}XW\varepsilon] = X^\intercal(XWX^\intercal)^{-1}XW\ Var(\varepsilon)\ W^\intercal X^\intercal(XWX^\intercal)^{-1}X = ...$$

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

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

Then we have,

$$(n-p)E[\hat{\sigma^2}] = E||Y - X^\intercal\hat{\theta}||^2 = E||\hat{\varepsilon}||^2 = E[tr(\hat{\varepsilon}\hat{\varepsilon}^\intercal)] = E[tr((I_n - \Pi)YY^\intercal(I_n - \Pi))] = ...$$

$$= E[tr((I_n - \Pi)(X^\intercal\theta^* + \varepsilon)(X^\intercal\theta^* + \varepsilon)^\intercal(I_n - \Pi))] = E[tr((I_n - \Pi)\varepsilon\varepsilon^\intercal(I_n - \Pi))] = tr((I_n - \Pi)E[\varepsilon\varepsilon^\intercal]) = ...$$

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

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

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

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

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

Exercise 3.

Consider the linear regression model from exercise 1. Suppose, that the target of estimation is $\gamma^\intercal\theta$ for some determinate non-zero vector $\gamma \in R^p$. Find expression for the LSE of $\gamma^\intercal\theta$. 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 $\gamma^\intercal\hat{\theta}$ since $E[\gamma^\intercal\hat{\theta} - \gamma^\intercal\theta] = \gamma^\intercal E[(XWX^\intercal)^{-1}XWY] - \gamma^\intercal\theta = \gamma^\intercal\theta - \gamma^\intercal\theta = 0$.

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

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

$$E[\tilde{\theta}] = E[((XWX^\intercal)^{-1}XW + D)Y] = E[((XWX^\intercal)^{-1}XW + D)(X^\intercal\theta + \varepsilon)] = ...$$
$$= E[\theta] + E[DX^\intercal\theta] + E[(XWX^\intercal)^{-1}XW)\varepsilon] = \theta + DX^\intercal\theta + 0$$

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

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

$$Var(\tilde{\theta}) = Var[((XWX^\intercal)^{-1}XW + D)Y] =$$
$$= [((XWX^\intercal)^{-1}XW + D)]Var(Y)[((XWX^\intercal)^{-1}XW + D)^\intercal] = ...$$
$$= [((XWX^\intercal)^{-1}XW + D)]Var(X^\intercal\theta^* + \varepsilon)[(WX^\intercal(XWX^\intercal)^{-1} + D^\intercal)] = ...$$
$$= [((XWX^\intercal)^{-1}XW + D)]W^{-1}[(WX^\intercal(XWX^\intercal)^{-1} + D^\intercal)] = ...$$
$$= [((XWX^\intercal)^{-1}XW + D)][(X^\intercal(XWX^\intercal)^{-1} + W^{-1}D^\intercal)] = ...$$
$$= (XWX^\intercal)^{-1} + DW^{-1}D^\intercal + DX^\intercal(XWX^\intercal)^{-1} + (XWX^\intercal)^{-1}XD^\intercal = Var(\tilde{\theta})$$

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

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

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

We have then

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

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

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:

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

Take,

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

,

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

this implies A is an orthogonal matrix, and further,

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

Exercise 5.

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

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

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

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

$$\Pi y = \hat{y} = X^\intercal(XX^\intercal)^{-1}Xy = X^\intercal\frac{Xy}{XX^\intercal} = \frac{\sum_i^n x_i y_i}{\sum_i^n x_i^2}(x_1, x_2, ..., x_n)^\intercal = \frac{< X \cdot y >}{< y \cdot y >}X^\intercal = Proj_X(y)$$

(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 $(I_n - \Pi)^\intercal(I_n - \Pi) = I_n - 2\Pi + \Pi = I_n - \Pi$, we have,

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

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

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

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 $span\{u_{p-k+1}, u_{p-k+2}, ..., u_p\} \subseteq span\{u_1, ..., u_p\}$.

(b) $||PL2x|| \leq ||PL1x|| \; \forall x \in R^n$,

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

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

and 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) $PL2 \cdot PL1 = PL2$
We can denote $P_{L_1}(x) = \sum_{i=1}^{p}(x \cdot u_i)u_i = UU^\intercal x$, where matrix $U_{n\times p}$ consists of orthnormal vectors $[u_1, ..., u_p]$, and denote

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

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

$$P_{L_2}P_{L_1} = VV^\intercal UU^\intercal$$

Since the first $p - k$ column vectors of V and U are the same, and orthonormal, the inner product $V^\intercal U$ 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

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

Section 2.1

Exercise 8.

Let $X \sim N(0, I_n)$, $Q = X^\intercal X$. Suppose that Q is decomposed into the sum of two quadratic forms: $Q = Q1+Q2$, where $Qi = X^\intercal A_i X$, $i = 1, 2$ for some symmetric matrices $A1, A2$ with $rank(A1) = n1$ and $rank(A2) = n2$. Show that if $n1 + n2 = n$, then $Q1$ and $Q2$ are independent and $Q_i \sim \chi^2(n_i) for \; i = 1, 2$.

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

Since $A1$ is a symmetric matrix, we can diagonalize it, $A_1 = U^\intercal \Lambda U$. We know the rank of A_1 is n_1. This implies that $U^\intercal A_1 U = \Lambda = diag\{\Lambda_1, ..., \Lambda_{n_1}, ..., \Lambda_n\}$, 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 $X^\intercal X = Y^\intercal U^\intercal U Y = Y^\intercal I_n Y = Y^\intercal Y$. So $Q = X^\intercal X = Y^\intercal Y = \sum_{i=1}^n Y_i^2$.

We can write

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

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

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

,

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 $Q_2 = X^\intercal X = Y^\intercal U^\intercal A_2 U Y$, from above we have

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

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 $\Lambda_1 = \Lambda_2 = ... = \Lambda_{n_1} = 1$. This also implies $Q = Q1 + Q2 = \sum_{i=1+1}^{n_1} Y_i^2 + \sum_{i=n_1+1}^{n} Y_i^2$.

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 $\eta = H^\intercal \theta^*$, where $H \in R^{q \times p}$ is some non-zero matrix with $q \leq p$. Find an analogue of the quadratic form S2 (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 $\eta^* = H^\intercal \theta^*$, and $\hat{\eta} = H^\intercal \hat{\theta}$, we have,

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

and

$$Var(H^\intercal \hat{\theta}) = H^\intercal Var(\hat{\theta}) H = H^\intercal Var((XX^\intercal)^{-1} X (X^\intercal \theta^* + \varepsilon)) H = H^\intercal Var(\theta^* + (XX^\intercal)^{-1} X \varepsilon) H = ...$$

$$... = H^\intercal ((XX^\intercal)^{-1} X \sigma^2 I_n X^\intercal (XX^\intercal)^{-1} H = \sigma^2 H^\intercal (XX^\intercal)^{-1} H = \sigma^2 S = Var(H^\intercal \hat{\theta})$$

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

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

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

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

Exercise 10.

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

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

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

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

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

So we have

$$\hat{\theta} = (XX^\intercal)^{-1}XY = \frac{n}{det(XX^\intercal)} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} = (\hat{\theta}_1, \hat{\theta}_2)^\intercal = \; ...$$

$$... \; = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \bar{y}\sum_i x_i^2 - \bar{x}\sum_i x_i y_i \\ \sum_i x_i y_i - n\bar{y}\bar{x} \end{bmatrix} = (\hat{\theta}_1, \hat{\theta}_2)^\intercal = \hat{\theta}$$

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

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

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 $F(2, n-2)$.

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

We know that $\frac{(XX^\intercal)^{1/2}(\hat{\theta} - \theta^*)}{\sigma} \sim N(0, I_p)$. 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:

$$T_{1(n-2-1)} = \frac{\hat{\theta_1} - \theta_1^*}{\sqrt{\hat{\sigma^2}[(XX^\intercal)^{-1}]_{11}}} = \frac{\hat{\theta_1} - \theta_1^*}{\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\theta_1} - \hat{\theta_2} x_i)^2}{n-p} \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

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:

$$T_{2(n-3)} = \frac{\hat{\theta_2} - \theta_2^*}{\sqrt{\hat{\sigma^2}[(XX^\intercal)^{-1}]_{22}}} = \frac{\hat{\theta_2} - \theta_2^*}{\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\theta_1} - \hat{\theta_2} x_i)^2}{n-p} \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

With T_2 we can set up a % $100(1-\alpha)$ 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 $var(\hat{Y}) = var(X^\intercal(XX^\intercal)^{-1}XY) = var(X^\intercal(XX^\intercal)^{-1}X(X^\intercal\theta^* + \varepsilon)) = var(X^\intercal(XX^\intercal)^{-1}X\varepsilon) = \sigma^2 X^\intercal(XX^\intercal)^{-1}X$. 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:

$$T_{(n-3)} = \frac{\hat{y}_i - y_i}{\sqrt{\hat{\sigma^2} x_i^\intercal (XX^\intercal)^{-1} x_i}} = \frac{\hat{y}_i - y_i}{\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\theta_1} - \hat{\theta_2} x_i)^2}{n-2} x_i^\intercal (XX^\intercal)^{-1} x_i}}$$

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

$$y_i \pm T_{n-3,\alpha/2} \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\theta_1} - \hat{\theta_2} x_i)^2}{n-2} x_i^\intercal \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} x_i}$$