

# Midterm 1: Math 6266

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:

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

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

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,

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

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,

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

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  $(I_n - \Pi)X^\top = X^\top - \Pi X^\top = X^\top - X^\top (X X^\top)^{-1} X X^\top = X^\top - X^\top = 0$ .

Then we have,

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

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  $\text{tr}(I_n - \Pi) = \text{tr}(I_n) - \text{tr}(\Pi) = n - p$ .

### 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  $\Pi$  (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,

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

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

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

(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,

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

## Section 2.1

*Exercise 8.* Let  $X \sim N(0, I_n)$ ,  $Q = X^\top X$ . Suppose that  $Q$  is decomposed into the sum of two quadratic forms:  $Q = Q_1 + Q_2$ , where  $Q_i = X^\top A_i X$ ,  $i = 1, 2$  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  $\Lambda$  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  $\Lambda$  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 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  $\theta^*$ ,  $\theta_1^*$ ,  $\theta_2^*$ .

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

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

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

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

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  $\theta^*$ , 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  $\theta^*$ , such that,  $qF_\alpha$  denotes the  $\alpha^{th}$  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  $\frac{(XX^\top)^{1/2}(\hat{\theta} - \theta^*)}{\sigma} \sim N(0, I_p)$ . We can then set up confidence intervals for  $\theta_1^*$  and  $\theta_2^*$ .

For  $\theta_1^*$ , 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  $\theta_2^*$  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 %  $100(1 - \alpha)$  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  $\text{var}(\hat{Y}) = \text{var}(X^\top (XX^\top)^{-1} XY) = \text{var}(X^\top (XX^\top)^{-1} X(X^\top \theta^* + \epsilon)) = \text{var}(X^\top (XX^\top)^{-1} X \epsilon) = \sigma^2 X^\top (XX^\top)^{-1} X$ . Using the standard error for  $\hat{Y}$ , we can set 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 %  $100(1 - \alpha)$  confidence interval for  $y_i$  can be created:

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