

Homework 2: Linear Regression

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN

DUE 10/13/2020

GO GREEN. AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

In class we studied the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}$, under the *homoskedastic* assumption $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. In this homework you will derive the same results for the slightly more general *heteroskedastic* model where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^*)$. Each subproblem is worth 10 points.

Problem 2.1. Derive an expression for the coefficient vector $\boldsymbol{\theta}$ that minimizes the mean squared error, i.e.,

$$\arg \min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2.$$

Problem 2.2. Derive an expression for the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}^*$, i.e.,

$$\arg \max_{\boldsymbol{\theta}} \mathbb{P}(\mathbf{y}, \mathbf{X} | \boldsymbol{\theta}, \boldsymbol{\Sigma}^*)$$

Problem 2.3. What is the distribution of the MLE of $\boldsymbol{\theta}^*$?

Problem 2.4. Given a new sample with feature vector \mathbf{x} , what is the MLE of the response, \hat{y} ?

Problem 2.5. Given a new sample with feature vector \mathbf{x} , what is the distribution of the MLE \hat{y} ?

Problem 2.6. Derive an expression for the MLE of $\boldsymbol{\Sigma}^*$, i.e.,

$$\arg \max_{\boldsymbol{\Sigma}} \mathbb{P}(\mathbf{y}, \mathbf{X} | \boldsymbol{\theta}^*, \boldsymbol{\Sigma})$$

Problem 2.7. Consider the following vector \mathbf{y} , containing information about glucose level of four individuals, and the following data matrix \mathbf{X} containing information about height and weight of the corresponding individuals:

$$\mathbf{y} = \begin{bmatrix} 110 \\ 140 \\ 180 \\ 190 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 180 & 150 \\ 150 & 175 \\ 170 & 165 \\ 185 & 210 \end{bmatrix}.$$

Given these data

- (a) What are your maximum likelihood estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\theta}^*$?
- (b) Given a new sample with feature vector $\mathbf{x} = [175 \ 170]^T$, what is the maximum likelihood estimate of its response \hat{y} ?
- (c) Derive a 95% confidence interval for \hat{y} .
- (d) Would you conclude that height is a significant feature for this model? Why?
- (e) Would you conclude that weight is a significant feature for this model? Why?

$$y = x\theta + \epsilon \quad \epsilon \sim N(0, \Sigma)$$

$$1. \quad Q = \|y - x\theta\|_v^2$$

$$\frac{\partial Q}{\partial \theta} = -2x^T(y - x\theta)$$

$$\left. \frac{\partial Q}{\partial \theta} \right|_{\hat{\theta}} = 0 \quad \Leftrightarrow \quad \hat{\theta} = (x^T x)^{-1} x^T y$$

$$\frac{\partial Q}{\partial \Sigma} = 2x^T x \geq 0$$

$$\text{Then } \hat{\theta} = (x^T x)^{-1} x^T y = \underset{\theta}{\operatorname{argmin}} \|y - x\theta\|_v^2$$

$$2. \quad y \sim N_n(x\theta^*, \Sigma^*)$$

$$f_{Y|Y}(y) = (2\pi)^{-n/2} (\det(\Sigma^*))^{-1/2} \exp \left\{ -\frac{1}{2} (y - x\theta^*)^T \Sigma^{*-1} (y - x\theta^*) \right\}$$

$$l(\theta) = \log(f_{Y|Y}(y)) = -\frac{n}{2} \log(\frac{\pi}{2}) - \frac{1}{2} \log(\det(\Sigma^*)) - \frac{1}{2} (y - x\theta^*)^T (\Sigma^*)^{-1} (y - x\theta^*)$$

$$\frac{\partial l(\theta)}{\partial \theta} = x^T (\Sigma^*)^{-1} (y - x\theta)$$

$$\left. \frac{\partial l(\theta)}{\partial \theta} \right|_{\hat{\theta}} = 0 \quad \Leftrightarrow \quad \hat{\theta} = (x^T (\Sigma^*)^{-1} x)^{-1} x^T (\Sigma^*)^{-1} y$$

$$\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T} = -x^T (\Sigma^*)^{-1} x \leq 0$$

$$\text{Then } \hat{\theta} = (x^T (\Sigma^*)^{-1} x)^{-1} x^T (\Sigma^*)^{-1} y \text{ is MLE}$$

$$3. \quad \text{Since } y \sim N_n(x\theta^*, \Sigma^*)$$

$\hat{\theta}$ is linear multiplication of y , is also multivariate normal distribution

$$E(\hat{\theta}) = (x^T (\Sigma^*)^{-1} x)^{-1} x^T (\Sigma^*)^{-1} x\theta^* = \theta$$

$$\begin{aligned} \operatorname{Var}(\hat{\theta}) &= (x^T (\Sigma^*)^{-1} x)^{-1} x^T (\Sigma^*)^{-1} \Sigma^* (\Sigma^*)^{-1} x (x^T (\Sigma^*)^{-1} x)^{-1} \\ &= (x^T (\Sigma^*)^{-1} x)^{-1} \end{aligned}$$

$$4. \text{ Since } \hat{y} = x_0 \hat{\theta} = g(\hat{\theta})$$

where g is a surjective function.

Since MLE is invariant through surjective function.

$$\hat{y}_{MLE} \text{ is } \hat{x}_0 \hat{\theta}_{MLE} = x_0 (x^T (\Sigma)^{-1} x)^{-1} x^T (\Sigma^{-1})^{-1} y$$

$$5. \text{ Use } \hat{y} \text{ to denote } \hat{y}_{MLE}$$

Since \hat{y} is linear multiplication of y . we have.

\hat{y} is multivariate normal distribution

$$\hat{y} \sim N_n(x_0, x_0 (x^T (\Sigma)^{-1} x)^{-1} x^T)$$

Problem 2.6. Derive an expression for the MLE of Σ^* , i.e.,

$$\arg \max_{\Sigma} \mathbb{P}(y, X | \theta^*, \Sigma)$$

$$f(y|\Sigma) = (2\pi)^{-\frac{n}{2}} (\det(\Sigma))^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y - X\theta^*)^\top \Sigma^{-1} (y - X\theta^*) \right\}$$

$$V(\Sigma) \rightarrow \log(f(y|\Sigma)) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(\det(\Sigma)) - \frac{1}{2} (y - X\theta^*)^\top \Sigma^{-1} (y - X\theta^*)$$

$$= -\frac{n}{2} \log(2\pi) + \frac{1}{2} \log(\det(\Sigma^{-1})) - \frac{1}{2} (y - X\theta^*)^\top \Sigma^{-1} (y - X\theta^*)$$

Since matrix is one-to-one mapping, then find MLE of Σ is equivalent to finding MLE of Σ^{-1}

$$\begin{aligned} \frac{\partial l(\Sigma)}{\partial \Sigma^{-1}} &= \frac{1}{2 \det(\Sigma^{-1})} \cdot \text{adj}(\Sigma^{-1}) - \frac{1}{2} \frac{\partial}{\partial \Sigma^{-1}} \left\{ \text{tr}((y - X\theta^*)(y - X\theta^*)^\top \Sigma^{-1}) \right\} \\ &= \frac{1}{2} \Sigma - \frac{1}{2} (X\theta^*)(X\theta^*)^\top = \frac{1}{2} [\Sigma - (y - X\theta^*)(y - X\theta^*)^\top] \\ \frac{\partial l(\Sigma)}{\partial \Sigma^{-1}} \Big|_{\hat{\Sigma}} &= 0 \quad \Leftrightarrow \quad \hat{\Sigma} = (y - X\theta^*)(y - X\theta^*)^\top \end{aligned}$$

(Claim: $f(A) = \log(\det(A))$ is a concave function of A for $\forall A > 0$ (A is positive definite matrix)

pf: let $g(t) = f(A + tB)$. $A, B > 0$

$$t \in \{t \in \mathbb{R} : A + tB > 0\}$$

$$g(t) = \log(\det(A + tB)) = \log \left(\det \left(A^{\frac{1}{2}} (I + tA^{-\frac{1}{2}} B A^{-\frac{1}{2}}) A^{\frac{1}{2}} \right) \right)$$

$$= \log \left\{ \det \left(I + tA^{-\frac{1}{2}} B A^{-\frac{1}{2}} \right) \det(A) \right\}$$

$$= \log \left\{ \det \left(I + tA^{-\frac{1}{2}} B A^{-\frac{1}{2}} \right) \right\} + \log \left\{ \det(A) \right\}$$

$$\text{let } D = A^{-\frac{1}{2}} B A^{-\frac{1}{2}}$$

Consider eigen decomposition of D . $D = Q \Lambda Q^T$

where D is orthogonal matrix. $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$

$$\begin{aligned} g(t) &= \log \left\{ \det(I + tQ\Lambda Q^T) \right\} + \log (\det(B)) \\ &= \log \left\{ \det(Q(I + t\Lambda)Q^T) \right\} + \log \left\{ \det(B) \right\} \\ &= \log \left\{ \det(I + t\Lambda) \right\} + \log (\det(B)) \\ &= \log \left\{ \prod_{i=1}^n (1 + t\lambda_i) \right\} + \log \left\{ \det(B) \right\} \\ &= \sum_{i=1}^n \log(1 + t\lambda_i) + \log(\det(B)) \end{aligned}$$

$$g'(t) = \sum_{i=1}^n \frac{\lambda_i}{1 + t\lambda_i} \quad g''(t) = \sum_{i=1}^n \frac{-\lambda_i^2}{(1 + t\lambda_i)^2} < 0$$

then $g(t)$ is concave function of t

then $f(B)$ is concave function of f \square

$$L(\Sigma) = -\frac{n}{2} \log(z_n) + \frac{1}{2} \log(\det(\Sigma^{-1})) - \frac{1}{2} (\mathbf{y} - \mathbf{x}\boldsymbol{\theta}^*)^\top \Sigma^{-1} (\mathbf{y} - \mathbf{x}\boldsymbol{\theta}^*)$$

Since $(\mathbf{y} - \mathbf{x}\boldsymbol{\theta}^*)^\top \Sigma^{-1} (\mathbf{y} - \mathbf{x}\boldsymbol{\theta}^*)$ is linear w.r.t. Σ^{-1} .

then $L(\Sigma)$ is concave w.r.t Σ^{-1}

Then $\frac{1}{2} = (\mathbf{y} - \mathbf{x}\boldsymbol{\theta}^*)(\mathbf{y} - \mathbf{x}\boldsymbol{\theta}^*)^\top$ is MLE of Σ^* \square

Problem 2.7. Consider the following vector \mathbf{y} , containing information about glucose level of four individuals, and the following data matrix \mathbf{X} containing information about height and weight of the corresponding individuals:

$$\mathbf{y} = \begin{bmatrix} 110 \\ 140 \\ 180 \\ 190 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 180 & 150 \\ 150 & 175 \\ 170 & 165 \\ 185 & 210 \end{bmatrix}.$$

Given these data

- (a) What are your maximum likelihood estimates of Σ and θ^* ?
- (b) Given a new sample with feature vector $\mathbf{x} = [175 \ 170]^T$, what is the maximum likelihood estimate of its response \hat{y} ?
- (c) Derive a 95% confidence interval for \hat{y} .
- (d) Would you conclude that height is a significant feature for this model? Why?
- (e) Would you conclude that weight is a significant feature for this model? Why?

Some computation solution is in notebook, please see appendix pages.

$$(a) \hat{\theta} = (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1} \mathbf{y}, \quad \hat{\Sigma} = (\mathbf{y} - \mathbf{X}\theta^*)(\mathbf{y} - \mathbf{X}\theta^*)^T$$

I use iteration to compute. $\theta = \begin{pmatrix} \text{intercept} \\ \text{height} \\ \text{weight} \end{pmatrix}$

$$\hat{\theta} = \begin{bmatrix} -50.46 \\ 0.12 \\ 1.06 \end{bmatrix} \quad \hat{\Sigma} = \begin{bmatrix} 374.0 & 341.6 & -668.5 & -26.5 \\ 341.6 & 311.9 & -610.6 & -24.2 \\ -668.5 & -610.6 & 1195.1 & 47.3 \\ -26.5 & -24.2 & 47.3 & 1.88 \end{bmatrix}$$

$$(b) \mathbf{x}_{\text{new}} = [175, 170]^T$$

$$\hat{y} = \mathbf{x}_{\text{new}} \cdot \hat{\theta} = 150.13$$

$$(c) \hat{y} \sim N_n(\mathbf{x}\theta, \mathbf{x}(\mathbf{x}^\top \Sigma^{-1} \mathbf{x})^{-1} \mathbf{x}^\top)$$

$$y^* = y_{\text{true}} = \mathbf{x}_0 \theta$$

$$\text{let } \mathbf{x}_0 = \mathbf{x}_{\text{new}} = [175, 170]^T$$

we have:

$$\frac{\hat{y} - x_0\theta}{\sqrt{x_0(x^T(\Sigma')^{-1}x)^{-1}x_0^T}} \sim N(0,1)$$

Since $\hat{\Sigma} \xrightarrow{\text{a.s.}} \Sigma'$ almost surely
according Strong Law of Large number.

$$\frac{x_0(x^T(\hat{\Sigma})^{-1}x)^{-1}x_0^T}{x_0(x^T(\Sigma')^{-1}x)^{-1}x_0^T} \xrightarrow{P} 1 \quad \text{in probability}$$

According to Slutsky theorem:

$$z = \frac{\hat{y} - x_0}{\sqrt{x_0(x^T(\hat{\Sigma})^{-1}x)^{-1}x_0^T}} = \frac{\hat{y} - x_0}{\sqrt{x_0(x^T(\Sigma')^{-1}x)^{-1}x_0^T}} \cdot \frac{x_0(x^T(\hat{\Sigma})^{-1}x)^{-1}x_0^T}{x_0(x^T(\Sigma')^{-1}x)^{-1}x_0^T}$$

$$\xrightarrow{d} N(0,1)$$

$$P(-z_{\alpha/2} \leq z \leq z_{\alpha/2}) = 1 - \alpha \quad \alpha = 0.05$$

$z_{\alpha/2}$ is top $\alpha/2$ quantile of $N(0,1)$. $z_{\alpha/2} = 1.96$

$$\hat{y} \pm \sqrt{x_0(x^T(\hat{\Sigma})^{-1}x)^{-1}x_0^T} \cdot z_{\alpha/2}$$

$$\sqrt{x_0(x^T(\hat{\Sigma})^{-1}x)^{-1}x_0^T} \cdot z_{\alpha/2} = 2.38 \times 10^{-7}$$

$$\hat{y} = 150.13$$

$$x_0 \in [150.13 - 2.38 \times 10^{-7}, 150.13 + 2.38 \times 10^{-7}]$$

$$(d) \quad \theta = \begin{pmatrix} \text{intercept} \\ \text{height} \\ \text{weight} \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \quad \hat{\theta} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\theta}_3 \end{pmatrix}$$

Since $\hat{\theta} \sim N(\theta, (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1})$

$$V_{\hat{\theta}} = (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \quad \hat{V}_{\hat{\theta}} = C \times 10^{-9}$$

$$H_0: \theta_2 = 0 \iff H_1: \theta_2 \neq 0$$

$$\left(\frac{\hat{\theta}_2}{\hat{V}_{\hat{\theta}}} \right)^2 \sim \chi^2_1$$

p-value > 0.05. height is not significant

$$(e) \quad V_3 = (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})_{33} \quad \hat{V}_3 = C \times 10^{-9}$$

$$H_0: \theta_3 = 0 \iff H_1: \theta_3 \neq 0$$

$$\left(\frac{\hat{\theta}_3}{\hat{V}_{\hat{\theta}}} \right)^2 \sim \chi^2_1$$

p-value > 0.05 weight is not significant

HW 2.7

```
In [68]: import numpy as np  
import pandas as pd
```

```
In [109]: y = np.array([110, 140, 180, 190])  
X = np.array([1, 180, 150, 1, 150, 175, 1, 170, 165, 1, 185, 210]).reshape(4,-1)  
X
```

```
Out[109]: array([[ 1, 180, 150],  
[ 1, 150, 175],  
[ 1, 170, 165],  
[ 1, 185, 210]])
```

(a)

```
In [110]: def mle_ite(Nsim, tol):  
    i = 0  
    theta = np.linalg.inv(X.T.dot(X)).dot(X.T).dot(y)  
    while i < Nsim:  
        sigma = (y - X.dot(theta)).reshape(-1,1).dot((y - X.dot(theta))  
        .reshape(1,-1))  
        thetal = np.linalg.inv(X.T.dot(np.linalg.inv(sigma)).dot(X)).d  
ot(X.T).dot(np.linalg.inv(sigma)).dot(y)  
        i += 1  
  
        if np.square((thetal - theta)).sum() <= tol:  
            print('Total %d' %i , 'iteration')  
            return thetal,sigma  
        theta = thetal  
  
    print('Total %d' %i , 'iteration')  
    return theta,sigma
```

```
In [111]: theta, sigma = mle_ite(1000,0.01)
```

Total 1 iteration

```
In [112]: theta, sigma
```

```
Out[112]: (array([-50.45722366,  0.11595569,  1.06057035]),  
 array([[ 380.26378754,  244.45529199, -697.15027716,  72.43119763]  
 ,  
       [ 244.45529199,  157.14983056, -448.16803531,  46.56291276]  
 ,  
       [-697.15027716, -448.16803531, 1278.10884145, -132.79052898]  
 ,  
       [ 72.43119763,  46.56291276, -132.79052898,  13.7964186 ]  
 )))
```

(b)

```
In [113]: x = np.array([1, 175, 170])
```

```
In [114]: x.dot(theta)
```

```
Out[114]: 150.13198208812736
```

(c)

```
In [115]: A = np.linalg.inv(X.T.dot(np.linalg.inv(sigma)).dot(X))  
A
```

```
Out[115]: array([[ 9.92008795e-15, -1.75058852e-16,  1.25555732e-16],  
                  [-2.29611375e-16,  2.31576551e-17, -1.94100457e-17],  
                  [ 1.73129595e-16, -1.95450278e-17,  1.64225156e-17]])
```

```
In [120]: ME = np.sqrt(x.dot(A).dot(x))*1.96  
ME
```

```
Out[120]: 2.3828115169069767e-07
```

```
In [119]: x.dot(theta) - ME
```

```
Out[119]: 150.1319818498462
```

(d)

```
In [122]: D = X.T.dot(np.linalg.inv(sigma)).dot(X)
D
```

```
Out[122]: array([-3.80643042e+14, -1.70978760e+17, -1.99172509e+17,
                   [-1.66683774e+17, -5.73462784e+19, -6.65041706e+19],
                   [-1.94363549e+17, -6.64473793e+19, -7.69884006e+19]])
```