统计机器学习 课后作业2

陈劭涵 17300180049

October 10, 2020

1 问题1

解:

For OLS estimator:

$RSS(\beta) = (y - X\beta)^T(y - X\beta)$

Thus we have:

$\frac{\partial RSS(\beta)}{\partial \beta} = 0 \iff \hat{\beta} = (X^TX)^{-1}(X^Ty)$

For MLE estimator:

$\because L(\beta) = \prod_{x_i} f(x, \beta)$

$\therefore L(\beta) = \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N} exp(-\frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2})$

Differentiating with respect to β, and we have:

$\frac{\partial L(\beta)}{\partial \beta} = 0 \iff \frac{\partial exp(-\frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2})}{\partial \beta} = 0$

$\because \frac{\partial exp(-\frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2})}{\partial \beta} = 0 \iff \frac{\partial (y-X\beta)^T(y-X\beta)}{\partial \beta} = 0$

$\therefore \hat{\beta} = (X^TX)^{-1}(X^Ty)$

In conclusion: OLS estimator $\iff$ MLE

2 问题2

解:

We already know that: $\hat{\beta} = \beta + (X^T X)^{-1} X^T \epsilon$

and $Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

Assume $\tilde{\beta}$ is any linear unbiased estimator without variance limits:

Denote $\tilde{\beta}$ as: $\tilde{\beta} = Cy$

where C is a $p \times N$ matrix

$\because \tilde{\beta}$ is unbiased

$\therefore E(Cy) = E(CX\beta + C\epsilon) = \beta$

$\therefore CX = I$ and $\tilde{\beta} = \beta + C\epsilon$

So the covariance matrix of $\tilde{\beta}$ is:

$Var(\tilde{\beta}) = \sigma^2 CC^T$

Now denote D as: $D \triangleq C - (X^T X)^{-1} X^T$, $D \neq 0$

Then we can denote $\beta^\star$ as $\beta^\star \triangleq \tilde{\beta} - \hat{\beta} = Dy$

$Var(\beta^\star) = D\Sigma_y D^T = \sigma^2 DD^T$

So DD^T is nonnegative definite matrix

Then we have:

$Var(\tilde{\beta}) = \sigma^2((D + (X^T X)^{-1} X^T)(D + (X^T X)^{-1} X^T)^T)$

$= \sigma^2((D + (X^T X)^{-1} X^T)(D^T + X(X^T X)^{-1}))$

$\because CX = I = DX + (X^T X)^{-1}(X^T X) = DX + I$

$\therefore DX = 0$

$\therefore Var(\tilde{\beta}) = \sigma^2(DD^T + (X^T X)^{-1})$

$= \sigma^2 DD^T + \sigma^2 (X^T X)^{-1}$

$= Var(\hat{\beta}) + \sigma^2 DD^T$

from the above we know that DD^T is nonnegative definite matrix

$\therefore Var(\tilde{\beta}) \geq Var(\hat{\beta})$

In conclusion, $\hat{\beta}$ has the smallest variance over all $\tilde{\beta}$

Therefore, Gauss-Markov Theorem is proved

3 问题3

解:

$\because \hat{\sigma}^2 = \frac{1}{N-p}(y - X\hat{\beta})^T(y - X\hat{\beta}) = \frac{1}{N-p}(y - \hat{y})^T(y - \hat{y})$

$\therefore E(\hat{\sigma}^2) = E(\frac{1}{N-p}(y - \hat{y})^T(y - \hat{y})) = \frac{1}{N-p}E((y - \hat{y})^T(y - \hat{y}))$

We focus on $E((y - \hat{y})^T(y - \hat{y}))$

In class we have mentioned:

$\hat{y} = X\hat{\beta} = X(X^TX)^{-1}(X^TY),\ let\ H \triangleq X(X^TX)^{-1}X^T$

Then $\hat{y} = Hy,\ y - \hat{y} = (I - H)y \Rightarrow X^T(y - \hat{y}) = 0$

Remark that I here refers to $I_{N\times N}$

We rewrite $I_{N\times N}$ as I_N

Take this equation into $(y - \hat{y})^T(y - \hat{y})$

$(y - \hat{y})^T(y - \hat{y}) = (X\beta + \epsilon - X\hat{\beta})^T(y - \hat{y}) = \epsilon^T(y - \hat{y})$

$\therefore (y - \hat{y})^T(y - \hat{y}) = \epsilon^T(y - \hat{y}) = \epsilon^T(I_N - H)y = \epsilon^T(I_N - H)(X\beta + \epsilon)$

$\because HX\beta = X\beta$

$\therefore (y - \hat{y})^T(y - \hat{y}) = \epsilon^T(I_N - H)\epsilon$

$\therefore E((y - \hat{y})^T(y - \hat{y})) = E(\epsilon^T(I_N - H)\epsilon)$

$\because \epsilon^T(I_N - H)\epsilon\ is\ a\ scalar$

$\therefore \epsilon^T(I_N - H)\epsilon = tr(\epsilon^T(I_N - H)\epsilon)$

According to the properties of the trace, we have:

$E(tr(\epsilon^T(I_N-H)\epsilon))=E(tr(\epsilon\epsilon^T(I_N-H)))=tr(E(\epsilon\epsilon^T(I_N-H)))$

$\because I_N-H\ is\ fixed$

$\therefore tr(E(\epsilon\epsilon^T(I_N-H)))=tr(E(\epsilon\epsilon^T)(I_N-H))=tr(\sigma^2(I_N-H))=\sigma^2tr(I_N-H)$

Note that $H\triangleq X(X^TX)^{-1}X^T$, and $rank(X)=p$

$\therefore tr(I_N-H)=tr(I_N-X(X^TX)^{-1}X^T)=tr(I_N-I_p)=N-p$

$\therefore E((y-X\hat{\beta})^T(y-X\hat{\beta}))=E((y-\hat{y})^T(y-\hat{y}))=(N-p)\sigma^2$

$\iff E(\hat{\sigma}^2)=\sigma^2$

4 问题4

解:

(1)We already know that: $\hat{\beta}=(X^TX)^{-1}X^Ty$

and: $y=X\beta+\epsilon$

$\therefore \hat{\beta}=\beta+(X^TX)^{-1}X^T\epsilon$

according to conditions β is fixed and $\epsilon\sim N(0,\sigma^2)$

$\therefore \hat{\beta}\ follows\ normal\ distribution\ and\ E(\hat{\beta})=\beta$

according to conditions A3-A5:

$\because Cov(\hat{\beta})=Cov((X^TX)^{-1}X^T\epsilon)=Cov(Z\epsilon)=ZCov(\epsilon)Z^T$

$=Z\sigma^2IZ^T=\sigma^2ZZ^T=\sigma^2(X^TX)^{-1}$

$\therefore Var(\hat{\beta})=Cov(\hat{\beta})=\sigma^2(X^TX)^{-1}$

$\therefore \hat{\beta}\sim N(\beta,\sigma^2(X^TX)^{-1})$

(2)In Problem 3, we have proved that:

$(y-\hat{y})^T(y-\hat{y})=\epsilon^T(I_N-H)\epsilon$

Note that $(N-p)\hat{\sigma}^2 = (y-\hat{y})^T(y-\hat{y})$

Therefore $(N-p)\hat{\sigma}^2 \sim \sigma^2\chi^2_{N-p} \iff \epsilon^T(I_N-H)\epsilon \sim \sigma^2\chi^2_{N-p}$

Now we prove that $\epsilon^T(I_N-H)\epsilon \sim \sigma^2\chi^2_{N-p}$:

$\because (I_N-H)(I_N-H) = I_N-2H+H^2 = I_N-2X(X^TX)^{-1}X^T+X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T$

$= I_N - X(X^TX)^{-1}X^T = I_N - H$

$\therefore (I_N-H)(I_N-H) = I_N - H$

$\therefore$ $I_N - H$ is a idempotent matrix with eigenvalue only 0 or 1

$\because I_N - H$ is also a symmetric matrix

$\therefore \exists$ orthogonal matrix Q and diagonal matrix Λ which satisfies:

$I_N - H = Q^T\Lambda Q$ and the diagonal element of Λ is 0 or 1

$\because$ From Problem 3 we proved that $tr(I_N-H) = N-p$

$\because$ Eigenvalue of $I_N - H$ and Λ is only 0 or 1

$\therefore$ There are $N-p$ of 1 and P of 0 in $I_N - H$' s N eigenvalues

and the same with Λ

$\therefore$ we have $\Lambda = \begin{pmatrix} I_{N-p} & 0 \\ 0 & 0 \end{pmatrix}$

From the above, we have:

$\epsilon^T(I_N-H)\epsilon = \epsilon^TQ^T\Lambda Q\epsilon$

$\epsilon^TQ^T\Lambda Q\epsilon = \sigma^2\frac{\epsilon^T}{\sigma}Q^T\Lambda Q\frac{\epsilon}{\sigma}$

We know that $\epsilon_i \sim N(0,\sigma^2)$

$\therefore \frac{\epsilon}{\sigma} \sim N(0, I_N)$

Because $N(0, I_N)$ keeps still under orthogonal transformation

$\therefore Q\frac{\epsilon}{\sigma} \sim N(0, I_N)$

Denote $Q\frac{\epsilon}{\sigma}$ as M

Then $\epsilon^T Q^T \Lambda Q \epsilon = \sigma^2 M^T \Lambda M = \sigma^2 M^T \begin{pmatrix} I_{N-p} & 0 \\ 0 & 0 \end{pmatrix} M$

$= \sigma^2 \sum_{i=1}^{N-p} M_i^2$

$\because$ each $M_i \sim N(0,1)$

$\therefore \sum_{i=1}^{N-p} M_i^2 \sim \chi^2_{N-p}$

$\therefore (N-p)\hat{\sigma}^2 = \epsilon^T (I_N - H)\epsilon = \epsilon^T Q^T \Lambda Q \epsilon = \sigma^2 \sum_{i=1}^{N-p} M_i^2 \sim \sigma^2 \chi^2_{N-p}$

5 问题5

解:

$log(y) = x^T\beta + \epsilon$

$\Rightarrow y = e^{x^T\beta+\epsilon}$

$E(y) = E(e^{x^T\beta+\epsilon}) = e^{x^T\beta}E(e^{\epsilon})$

$= e^{x^T\beta}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-t^2}{2\sigma^2}} e^t dt = e^{x^T\beta}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-t^2+2\sigma^2 t}{2\sigma^2}} dt = e^{x^T\beta}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(t-\sigma^2)^2+\sigma^4}{2\sigma^2}} dt$

$= e^{x^T\beta}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(t-\sigma^2)^2+\sigma^4}{2\sigma^2}} d(t-\sigma^2) = e^{x^T\beta} \cdot e^{\frac{\sigma^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(t-\sigma^2)^2}{2\sigma^2}} d(t-\sigma^2)$

$= e^{x^T\beta} \cdot e^{\frac{\sigma^2}{2}} = e^{x^T\beta+\frac{\sigma^2}{2}}$

6 问题6

解:

$TSS = \sum_i (y_i - \bar{y})^2 = \sum_i \left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right]^2$

$= \sum_i (y_i - \hat{y}_i)^2 + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_i (\hat{y}_i - \bar{y})^2$

Now we prove that $2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

$\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

$\iff \sum_i(y_i - \hat{y}_i)\hat{y}_i - \sum_i(y_i - \hat{y}_i)\bar{y} = 0$

Take $X, \hat{\beta}$ into the equation and tranform it into vectors

$\iff (y - \hat{y}^T)X\hat{\beta} - \bar{y} \cdot I_{N\times 1}(y - \hat{y}) = 0$

According to the previous analysis, we know that:

Because $\frac{\partial RSS(\beta)}{\partial \beta} = 0 \Rightarrow X^T(y - \hat{y}) = 0$ and $\bar{y} \cdot I_{N\times 1}(y - \hat{y}) = 0$

$\iff 0 - 0 = 0$

$\therefore \sum_i(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$ is true

$\therefore TSS = \sum_i(y_i - \hat{y}_i)^2 + \sum_i(\hat{y}_i - \bar{y})^2$

$\iff TSS = ESS + RSS$