作业2

(1) Prove that the OLS estimator $\widehat{\beta}$ is the same as the maximum likelihood estimator.

Proof: OLS estimation $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$

Under the assumption that ε_i's are normal, i.e., $\varepsilon \sim N(0, \sigma^2 I)$, and suppose $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$, we have the likelihood function of β:

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\underbrace{(y_i - x_i^T\beta)^2}_{\varepsilon_i}\right\}$$

and log-likelihood of β:

$$\ell(\beta) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$

To maximize $\ell(\beta)$, we need to minimize $\sum (y_i - x_i^T\beta)^2$, then the MLE of β is:

$$\tilde{\beta}_{MLE} = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 = \hat{\beta}_{OLS}$$

Therefore, the OLS estimation of β is the same as the maxium likelihood estimation of β. $\square$

(2) Prove the Gauss-Markov Theorem.

Proof: Assume that $E(\varepsilon)=0$ 无偏 and $Cov(\varepsilon)=\sigma^2 I_n$, 同方差

OLS estimate is $\hat{\beta}$ and the other linear unbiased estimate is $\tilde{\beta} = B\cdot y$. , $B \in \mathbb{R}^{p\times n}$

Since $E\tilde{\beta} = \beta$ for all possible β's (definition of "unbias")

$\Rightarrow \quad E\tilde{\beta} = E(By) = E(B(X\beta+\varepsilon)) = BX\beta = \beta$

$\Rightarrow \quad BX = I_p$ and $\tilde{\beta} = B(X\beta+\varepsilon) = \beta + B\varepsilon$

On the other hand:

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

$$\begin{aligned}
Var(\tilde{\beta}) &= Var(B\varepsilon) \\
&= Var((X^TX)^{-1}X^T\varepsilon + (B-(X^TX)^{-1}X^T)\varepsilon) \\
&= Var(\hat{\beta}) + Var((B-(X^TX)^{-1}X^T)\varepsilon) \\
&\quad + C + C^T
\end{aligned}$$

where: $C = E[((X^TX^{-1})^{-1}X^T\varepsilon - 0)\cdot((B-(X^TX)^{-1}X^T)\varepsilon - 0)^T]$

$$= (X^TX^{-1})^{-1}X^TE(\varepsilon\varepsilon^T)(B^T - X(X^TX)^{-T})$$

$$= \sigma^2 (X^TX^{-1})X^T(B^T - X(X^TX)^{-1}) = 0$$

For $\forall \eta \in \mathbb{R}^p$, $\|\eta\|_2 = 1$:

$$Var(\eta^T\tilde{\beta}) - Var(\eta^T\hat{\beta})$$

$$= \eta^T \left[Var(\tilde{\beta}) - Var(\hat{\beta}) \right] \eta$$

$$= \eta^T \left[Var\left((B - (X^T X)^{-1} X^T) \varepsilon \right) \right] \eta$$

$$\geq 0$$

$$\Rightarrow \quad Var(\eta^T \tilde{\beta}) \geq Var(\eta^T \hat{\beta})$$

$\Rightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T y$ is the "Blue" of β $\square$

(3) Prove $E(\widehat{\sigma}^2) = \sigma^2$.

Proof: $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = P_X y$

where $P_X = X(X^T X)^{-1} X^T$

P is a projecting matrix s.t.

$P_X X = X$ and $P_X^2 = P_X$ and $P_X^T = P_X$

$$\hat{\sigma}^2 = \frac{1}{N-p} \| y - \hat{y} \|^2$$

$$= \frac{1}{N-p} \| y - P_X y \|^2$$

$$= \frac{1}{N-p} \| (I_N - P_X) y \|^2$$

$$= \frac{1}{N-p} \, y^T (I_N - P_X)^T (I_N - P_X) \, y$$

$$= \frac{1}{N-p} \, y^T (I_N - P_X) \, y$$

$$= \frac{1}{N-p} (X\beta + \varepsilon)^T (I_N - P_X) (X\beta + \varepsilon)$$

$$= \frac{1}{N-p} \left[\beta^T X^T (I_N - P_X) X\beta + 2 \varepsilon^T (I_N - P_X) X\beta + \varepsilon^T (I_N - P_X) \varepsilon \right]$$

$$= \frac{1}{N-p} \varepsilon^T (I_N - P_X) \, \varepsilon$$

And $E[\hat{\sigma}^2] = E\left[\frac{1}{N-p} \varepsilon^T (I_N - P_X) \varepsilon\right]$

$$= \frac{1}{N-p} E\left[\operatorname{trace}\left((I_N - P_X) \varepsilon \varepsilon^T \right)\right]$$

$$= \frac{1}{N-p} \operatorname{trace}\left\{ E\left[(I_N - P_X) \varepsilon \varepsilon^T \right] \right\}$$

$$= \frac{1}{N-p} \operatorname{trace}\left\{ (I_N - P_X) \cdot E[\varepsilon \varepsilon^T] \right\}$$

$$= \frac{1}{N-p} \operatorname{trace}\left\{ (I_N - P_X) \, \sigma^2 \cdot I_N \right\}$$

$$= \frac{\sigma^2}{N-p} \operatorname{trace}\left\{ I_N - P_X \right\}$$

$$= \frac{\sigma^2}{n-p} (n - \mathrm{trace}(P_X))$$

$$= \frac{\sigma^2}{N-p} (n - \mathrm{trace}(X(X^T X)^{-1} X^T))$$

$$= \frac{\sigma^2}{N-p} (n - \mathrm{trace}((X^T X)^{-1} X^T X))$$

$$= \frac{\sigma^2}{N-p} (n - p)$$

$$= \sigma^2 \qquad \square$$

(4) Given conditions:

(A1) The relationship between response ($\mathbf{y}$) and covariates ($\mathbf{X}$) is linear;

(A2) $\mathbf{X}$ is a non-stochastic matrix and rank$(\mathbf{X}) = p$;

(A3) $E(\varepsilon) = \mathbf{0}$. This implies $E(\mathbf{y}) = \mathbf{X}\beta$;

(A4) $\mathrm{cov}(\varepsilon) = E(\varepsilon\varepsilon^\top) = \sigma^2 I_N$; (Homoscedasticity)

(A5) ε follows multivariate normal distribution $N(\mathbf{0}, \sigma^2 I_N)$ (Normality)

Prove the following results:

$$\widehat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}) \tag{0.1}$$

$$(N-p)\widehat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p} \tag{0.2}$$

Proof: $\varepsilon \sim N(0, \sigma^2 I_N)$, $\quad y = X\beta + \varepsilon$

$\Rightarrow y \sim N(X\beta, \sigma^2 I_N)$

$\hat{\beta} = (X^T X)^{-1} X^T y$, $\quad y \sim N(X\beta, \sigma^2 I_N)$

$\Rightarrow \hat{\beta} \sim N((X^T X)^{-1} X^T X\beta, (X^T X)^{-1} X^T \sigma^2 I_N X (X^T X)^{-1})$

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

$$\hat{\sigma}^2 = \frac{1}{N-p} \varepsilon^T (I_N - P_X) \varepsilon$$

(P_X follows the definition in the above problem.)

$$\Rightarrow (N-p)\hat{\sigma}^2 = \varepsilon^T (I_N - P_X) \varepsilon \qquad P_X = X(X^T X)^{-1} X^T$$

Assume the SVD of X is $X = U \Sigma V^T$

where $\text{rank}(\Sigma) = p$, $U U^T = I_N$, $V V^T = I_p$

$$\Rightarrow P_X = X (X^T X)^{-1} X^T$$

$$= U \Sigma V^T (V \Sigma^T \Sigma V^T)^{-1} V \Sigma^T U^T$$

$$= U \Sigma V^T V (\Sigma^T \Sigma)^{-1} V^T V \Sigma^T U^T$$

$$= U \Sigma (\Sigma^T \Sigma)^{-1} \Sigma^T U^T$$

$$\Rightarrow \varepsilon^T (I_N - P_X) \varepsilon$$

$$= \varepsilon^T (U U^T - U \Sigma (\Sigma^T \Sigma)^{-1} \Sigma^T U^T) \varepsilon$$

Let $\xi = U^T \varepsilon$, since $\varepsilon \sim N(0, \sigma^2 I_N)$

then $\xi \sim N(0, U^T \sigma^2 I_N U) = N(0, \sigma^2 I_N)$

$$\Rightarrow \xi^T (I_N - \Sigma (\Sigma^T \Sigma)^{-1} \Sigma^T) \xi$$

Since $\Sigma_{ij} = \begin{cases} \sigma_i & \text{if } i=j \text{ and } i \le p \\ 0 & \text{o.w} \end{cases}$

$$\Rightarrow \Sigma(\Sigma^T\Sigma)^{-1}\Sigma^T = \begin{bmatrix} I_p & \\ & 0_{N-p} \end{bmatrix}_{N\times N}$$

$$\Rightarrow \xi^T (I_N - \Sigma (\Sigma^T \Sigma)^{-1} \Sigma^T) \xi$$

$$= \sum_{i=N-p+1}^{N} \xi_i^2$$

Since $\xi_i \sim N(0, \sigma^2)$, we have:

$$\sum_{i=N-p+1}^{N} \xi_i^2 \sim \sigma^2 \chi^2_{N-p} .$$

That is: $(N-p)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p}$ $\square$

(5) Suppose y follows the log-linear regression relationship with $x \in \mathbb{R}^p$, i.e.,

$$\log(y) = x^\top \beta + \epsilon, \tag{0.3}$$

where ϵ follows normal distribution $N(0, \sigma^2)$. Please calculate $E(y)$.

Solution:

$$y = \exp\{x^T\beta + \varepsilon\}, \quad \varepsilon \sim N(0, \sigma^2)$$

$$\Rightarrow E[y] = E[e^{x^T\beta + \varepsilon}]$$
$$= E[e^{x^T\beta} \cdot e^{\varepsilon}]$$
$$= e^{x^T\beta} \cdot E[e^{\varepsilon}]$$

$$E[e^{\varepsilon}] = \int_{-\infty}^{+\infty} e^{t} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{\frac{-t^2}{2\sigma^2}\right\} dt$$

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(t-\sigma^2)^2+\sigma^4}{2\sigma^2}\right\} dt$$

$$= \exp\left\{\frac{\sigma^2}{2}\right\} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(t-\sigma^2)^2}{2\sigma^2}\right\} dt$$

$$= \exp\left\{\frac{\sigma^2}{2}\right\}$$

$$\Rightarrow E[y] = \exp\left\{x^T\beta + \frac{\sigma^2}{2}\right\} \qquad \square$$

(6) Define $\widehat{y}_i = x_i^\top \beta$. Let the intercept be included in the regression model. Define the total sum of squares (TSS) and explained sum of squares (ESS) as follows

$$\text{TSS} = \sum_i (y_i - \overline{y})^2, \quad \text{ESS} = \sum_i (\widehat{y}_i - \overline{y})^2.$$

Please prove:

$$\text{TSS} = \text{ESS} + \text{RSS}.$$

Proof: $RSS = \sum_i (y_i - \hat{y}_i)^2$

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

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

$$= RSS + ESS + 2R$$

Then we need to show that $R = 0$.

$$RSS = \sum_{i=1}^{N} (y_i - x_{i0}\beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2$$

Where $x_{i0}\beta_0$ is the intercept and $x_{i0} = 1$

From the definition of LSE,

$$\frac{\partial RSS}{\partial \beta_k} = 0, \quad \text{for } k = 0, 1, \cdots, p$$

When we take $k=0$, we have:

$$\frac{\partial RSS}{\partial \beta_0} = \sum_{i=1}^{N} (y_i - x_{i0}\beta_0 - \cdots - x_{ip}\beta_p) = \sum_{i=1}^{N} (y_i - \hat{y}_i) = 0$$

Then: $R = \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}_i)$

$$= (y - \hat{y})^T \cdot (\hat{y} - \bar{y})$$

$= (y-\hat{y})^T \cdot \hat{y} - (y-\hat{y})^T \bar{y}$

Since the residual $\varepsilon = y - \hat{y}$ is orthogonal to $\hat{y}$,
we have $(y-\hat{y})^T \hat{y} = 0$

And since $\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$, we have:

$(y-\hat{y})^T \bar{y} = 0$

Now we can conclude that $R = 0$ and

$TSS = ESS + RSS$

$\square$