cmi

Chennai Mathematical Institute

Regression and Classification

Sourish Das

Mid-sem Exam

3rd October 2023

Answer all 6 questions. Write briefly and to the point.
Total Time: 2 hours Total Marke: 30

1. "If the correlation between two predictors is high then the least square estimator of the linear regression model becomes unreliable." - Why? (3 points)

2. The Ridge estimator for the coefficients of the regression model is defined as

$X^T\beta$.

$$\hat{\beta}_{Ridge} = (X^TX + \lambda I)^{-1}X^Ty$$

$y = X\beta +$
$(y - X$

Show Ridge estimator is a biased estimator? (3 points)

3. If error structure, in linear models, follows $N(0, \sigma^2)$, then find the sampling distribution of the $\hat{\beta}_{Ridge}$. (3 points)

4. Why LASSO is effective feature selection tool than best-subset selection or forward selection process? (3 points)

5. Write down the following time-series model in linear model format,

$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2), \quad \mathbb{P}(y_0 = 0) = 1, \quad and \quad t = 1, 2, \cdots, T;$$

and find the OLS estimator for β_0 and β_1. (6 points)

6. Daily air quality measurements in New York during 1973 is available in `airquality` dataset available in `datasets` R-package. Following regression model was fitted

$$\texttt{Ozone} = \beta_0 + \beta_1\texttt{Solar.R} + \beta_2\texttt{Wind} + \beta_3\texttt{Temp} + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

where,

`Ozone` is Mean ozone in parts per billion at Roosevelt Island of NY City,

`Solar.R`: Solar radiation in Langleys in the frequency band 4000–7700 Angstroms at Central Park,

`Wind`: Average wind speed in miles per hour at LaGuardia Airport, and

`Temp`: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

$y = X\beta + \epsilon$.

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 & \cdots & x_p \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Following analysis using R is presented below:

```
Call:
lm(formula = Ozone ~ Solar.R + Wind + I(Wind^2) + Temp + I(Temp^2))

Residuals:
    Min      1Q  Median      3Q     Max
-48.017 -10.810  -4.144   8.120  80.125

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 291.09564  101.00727   2.882  0.00479 **
Solar.R       0.06593    0.02007   3.285  0.00139 **
Wind        -13.37647    2.30330  -5.808 6.83e-08 ***
I(Wind^2)     0.46372    0.10087   4.597 1.20e-05 ***
Temp         -6.34116    2.72014  -2.331  0.02165 *
I(Temp^2)     0.05104    0.01777   2.873  0.00492 **
---

Residual standard error: 18.27 on 105 degrees of freedom
Multiple R-squared:  0.7123,Adjusted R-squared:  0.6986
F-statistic: 51.99 on 5 and 105 DF,  p-value: < 2.2e-16
```

(i) Provide estimate of σ. (3 point)

(ii) If `Solar.R` = 185, `Wind` = 10 and `Temp` = 78, then compute expected `Ozone` level and 95% Confidence Interval of the `Ozone` level. (3 points)

(iii) Which predictor has strongest influence on `Ozone` level and why? (3 point)

(iv) What `Adjusted R-squared` explain with respect to model? (3 point)

$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ $B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$ $A+B = \begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix}.$

$= \dfrac{\begin{bmatrix} 4 & -3 \\ -2 & 1 \end{bmatrix}^T}{-2}$ $B^{-1} = \dfrac{\begin{bmatrix} 8 & -7 \\ -6 & 5 \end{bmatrix}^T}{-2}$ $(A+B)^{-1} = \dfrac{\begin{bmatrix} 12 & -8 \\ -10 & 6 \end{bmatrix}.}{-8}$

$= \dfrac{\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}}{-2}$ $= \dfrac{\begin{bmatrix} 8 & -6 \\ -7 & 5 \end{bmatrix}}{-2}$ $=$

$\dfrac{\text{adj } A}{\det A} + \dfrac{\text{adj } B}{\det B}$ $\dfrac{\text{adj }(A+B)}{\det(A+B)}$

$A^{-1} + B^{-1} = -\dfrac{1}{2} \begin{bmatrix} 12 & -8 \\ -10 & 6 \end{bmatrix}$

$= \begin{bmatrix} -6 & 4 \\ 5 & 3 \end{bmatrix}.$

$\det B$

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

$(X^TX)\beta = X^Ty.$

$\beta = \dfrac{X^Ty}{X^TX}$

$RSS \sim \sigma^2 \chi^2_{n-p}.$

$\dfrac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}.$

$Y = X\beta + \epsilon.$

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

$= (Y^T - \beta^TX^T)(Y - X\beta).$

$\sum ((y_i - x_i^T\beta_i)^2)$

$\dfrac{\partial RSS}{\partial \beta} = \partial\beta$

$= 2(y_i - x_i^T\beta_i).(-x_i^T).$

$Y = X\beta + \lambda\beta^T\beta.$