cmi

Chennai Mathematical Institute

Regression and Classification

Sourish Das

Mid-sem Exam

2nd October 2024

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

1. Show that the least squares method guarantees at least one solution. (3 points)

2. "When there is a high correlation between two predictors, the least squares estimator in a linear regression model becomes unstable and unreliable. - Why? (3 points)

3. (3 points)

   (a) The Ridge estimator for the coefficients of the regression model is defined as

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

   Show Ridge estimator is a biased estimator?

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

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. Twelve subjects were given oral doses of theophylline then serum concentrations were measured at 11 time points over the next 25 hours. The data is available in `Theoph` dataset available in `datasets` R-package. The datasets cosists of following variables:

   `Dose` dose of theophylline administered orally to the subject (mg/kg),

   `Time`: time since drug administration when the sample was drawn (hr), and

   `conc`: theophylline concentration in the sample (mg/L).

Following analysis using R is presented below:

```
Call:
lm(formula = log(conc + 1) ~ Time + I(Time^2) + Dose + I(Dose^2),
    data = Theoph)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.54738 -0.26046  0.06115  0.41075  0.93670 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.5127840  1.5402729   1.631    0.105    
Time         0.1141491  0.0228032   5.006 1.82e-06 ***
I(Time^2)   -0.0058816  0.0009534  -6.169 8.50e-09 ***
Dose        -0.5333977  0.6837919  -0.780    0.437    
I(Dose^2)    0.0629102  0.0750044   0.839    0.403    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.583 on 127 degrees of freedom
Multiple R-squared:  0.2647,	Adjusted R-squared:  0.2416 
F-statistic: 11.43 on 4 and 127 DF,  p-value: 5.892e-08
```

(i) Provide estimate of σ. (3 point)

(ii) If Dose = 4.0, Time = 1.25, then compute expected conc level and 95% Confidence Interval of the conc level. (3 points)

(iii) Which predictor has strongest influence on conc level and why? (3 points)

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