MATH 423/533 - ASSIGNMENT 3

To be handed in not later than 11:59pm, 20th November 2017.
Please submit your solutions as pdf via myCourses using the template provided.

For this assignment, you need to download a data set from the course website: this data set is

`http://www.math.mcgill.ca/yyang/regression/data/cigs.csv`

The data relate to a study of 25 cigarette brands: in the data set, for each brand,

- x_{i1} is the tar content (mg), denoted `TAR`;
- x_{i2} is the nicotine content (mg), denoted `NICOTINE`;
- x_{i3} is the weight (g), denoted `WEIGHT`;
- y_i is the amount of Carbon Monoxide (mg) produced in a standardized volume, denoted `CO`.

Regression models constructed to study the ability of the predictors to capture the variation in response are considered. The most complex model considered is the multiple regression model

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[Y_i|\mathbf{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

All models to be considered are nested within this one.

For your data set, compute and report the following quantities, all written in the notation from lectures.

(a) $\text{SS}_{\text{Res}}(\beta_0, \beta_1, \beta_2, \beta_3)$.

(b) $\text{SS}_{\text{Res}}(\beta_0, \beta_1, \beta_2)$.

10 Marks

(c) The F test statistic for comparing the two models.

10 Marks

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[Y_i|\mathbf{x}_i] &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \\ \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[Y_i|\mathbf{x}_i] &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}\end{aligned}$$

(d) The terms in the decomposition

10 Marks

$$\overline{\text{SS}}_{\text{R}}(\beta_1, \beta_2, \beta_3|\beta_0) = \overline{\text{SS}}_{\text{R}}(\beta_3|\beta_0) + \overline{\text{SS}}_{\text{R}}(\beta_2|\beta_0, \beta_3) + \overline{\text{SS}}_{\text{R}}(\beta_1|\beta_0, \beta_3, \beta_2)$$

(e) The terms in the decomposition

20 Marks

$$\overline{\text{SS}}_{\text{R}}(\beta_1, \beta_2|\beta_0) = \overline{\text{SS}}_{\text{R}}(\beta_1|\beta_0) + \overline{\text{SS}}_{\text{R}}(\beta_2|\beta_0, \beta_1)$$

(f) The F test statistic for comparing the two models

20 Marks

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[Y_i|\mathbf{x}_i] &= \beta_0 + \beta_1 x_{i1} \\ \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[Y_i|\mathbf{x}_i] &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\end{aligned}$$

if it is known that the predictor x_{i3} is not influential.

(g) The F test statistic for comparing the two models

15 Marks

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[Y_i|\mathbf{x}_i] &= \beta_0 \\ \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[Y_i|\mathbf{x}_i] &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\end{aligned}$$

if it is known that the predictor x_{i3} is not influential.

15 Marks

EXTRA QUESTION 1 FOR STUDENTS IN MATH 533

Suppose that principal component regression forms the derived input columns $\mathbf{z}_m = \mathbf{X}v_m$, where v_m's are the principle directions, and then regresses $\mathbf{y}$ on $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M$ for some $M \leq p$.

- Show that when the $\mathbf{z}_m$ are orthogonal, this regression is just a sum of univariate regressions:

$$\hat{\mathbf{y}}^{\text{pcr}}_{(M)} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m$$

  where $\hat{\theta}_m = \mathbf{z}_m^T\mathbf{y}/\mathbf{z}_m^T\mathbf{z}_m$.

  10 Marks

- Show that the $\mathbf{z}_m$ are each linear combinations of the original $\mathbf{x}_j$, we can express the solution in terms of coefficients of the $\mathbf{x}_j$.

$$\hat{\beta}^{\text{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m$$

  and show that $\hat{\beta}^{\text{pcr}}(p) = \hat{\beta}^{\text{ls}}$, where p is the number of variables in the model. 10 Marks

EXTRA QUESTION 2 FOR STUDENTS IN MATH 533

The following two information criteria are commonly used to compare regression models: if the log-likelihood of the data under model M is denoted ℓ_M, and p predictors are used to specify the model, then we have

- Akaike's Information Criterion (AIC):

$$\text{AIC}_M = -2\ell_M(\widehat{\beta}, \widehat{\sigma}_{\text{ML}}) + 2(p+1)$$

- Schwarz's Bayesian Information Criterion (BIC):

$$\text{BIC}_M = -2\ell_M(\widehat{\beta}, \widehat{\sigma}_{\text{ML}}) + (p+1)\log(n)$$

  where n is the sample size.

For a regression model under standard assumptions

$$\ell_M(\beta, \sigma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)$$

We will assume here, that unlike for the F-test procedure, these criteria can used to compare non-nested models. The function `AIC` in `R` can return either AIC and BIC values.

For the data in the main question, tabulate the AIC and BIC for all the models nested within the 'full' model, which includes all three predictors

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[Y_i|\mathbf{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

The model to be selected under AIC or BIC is the one for which the evaluated criterion is smallest. On the basis of your AIC and BIC computations, select the best regression model for the data. Assess, using the usual checks, whether the selected model is adequate.

20 Marks