

MATH 423/533 - ASSIGNMENT 3

*To be handed in not later than 10pm, 16th November 2016.
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/dstephens/Regression/Data/Assignment3/XXXXXXXXXX.csv>

where XXXXXXXXXX is your McGill ID number. 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) $SS_{\text{Res}}(\beta_0, \beta_1, \beta_2, \beta_3)$.

2 Marks

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

2 Marks

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

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

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

2 Marks

(d) The terms in the decomposition

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

4 Marks

(e) The terms in the decomposition

$$\overline{SS}_R(\beta_1, \beta_2|\beta_0) = \overline{SS}_R(\beta_1|\beta_0) + \overline{SS}_R(\beta_2|\beta_0, \beta_1)$$

4 Marks

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

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

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

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

3 Marks

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

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[Y_i|\mathbf{x}_i] = \beta_0$$

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

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

3 Marks

EXTRA QUESTION 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):

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

- Schwarz's Bayesian Information Criterion (BIC):

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

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 be 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.

5 Marks